

BALANCED REALIZATIONS FOR LINEAR SYSTEMS: A VARIATIONAL APPROACH*

U. HELMKE†

Abstract. This paper develops a new geometric approach to balanced realizations, which enables balanced realizations for arbitrary linear systems to be defined and studied. For an arbitrary (unitarily invariant) strictly plurisubharmonic function on the set of realizations of a given transfer function, the class of realizations is considered that minimizes the function. Based on results from invariant theory and complex analysis, a general theorem on the existence and uniqueness properties of such function-minimizing realizations is derived. If the function is the sum of the traces of the controllability and observability gramians, the usual class of balanced realizations is obtained. Other choices of functions yield different, new classes of function-minimizing realizations.

Key words. balanced realizations, linear systems, norm minimization, invariant theory

AMS(MOS) subject classifications. 93B20, 93B27

1. Introduction. Balanced realizations in the state space for asymptotically stable linear systems are introduced by Moore [16] and have quickly found widespread use in model reduction and approximation theory of linear systems. They are defined by the characterizing property that their controllability and observability gramians are equal and diagonal and an explicit construction is available based on the singular value decomposition (SVD) of the associated Hankel operators. In signal processing, balanced realizations and singular values already appear in the early work of Mullis and Roberts [17], Hwang [10]. For parametrizations of classes of linear systems by balanced realizations, we refer to Ober [21].

The current use of balanced realizations is mainly restricted to the class of asymptotically stable linear systems, while only few attempts have been made to widen the class of linear systems that can be balanced. Kenney and Hower [13] introduce balanced realizations for a certain generic class of transfer functions, which also contains unstable systems. A different type of (LQG) balanced realizations was introduced by Verriest [25] and further studied by Jonkheere and Silverman [11]. Their construction is based on the unique positive definite solution of the algebraic Riccati equation and thus works for arbitrary minimal systems (A, B, C) .

From a theoretical viewpoint, the problem of finding balanced realizations of linear systems is very similar to a classical problem in physics, namely, that of diagonalizing the inertia tensor of a rigid body; see Arnold [3]. It is therefore of interest to see whether it is also possible to analyze balanced realizations in the same way as physicists analyze momentum tensors, i.e., via coadjoint orbits and moment maps.

In this paper, we develop such a new geometric approach, which enables us to define and study balanced realizations for arbitrary, i.e., possibly unstable, transfer functions in a coherent and systematic framework. Our main tool is a recent result due to Kempf and Ness [12], who consider moment maps in the context of invariant theory. The starting point for our analysis are arbitrary (unitarily invariant) Hermitian norms on the set of realizations (A, B, C) of a given transfer function. All realizations that minimize a given norm are called *norm minimal*, and, based on the Kempf–Ness theorem, we derive a general existence and uniqueness theorem for such norm minimal realizations. The result is then generalized for the class of smooth, unitarily invariant,

* Received by the editors May 23, 1990; accepted for publication (in revised form) August 2, 1991.

† Department of Mathematics, University of Regensburg, 8400 Regensburg, Germany.

strictly plurisubharmonic functions, a special case being the plurisubharmonic function defined as the sum of traces of the controllability and observability gramians. If the function is the sum of traces of the controllability and observability gramians, the class of balanced realizations introduced by Moore is obtained. Other choices of functions yield different, new classes of function-minimizing realizations.

In § 2 the Kempf–Ness theorem (Theorem 2.1) is reviewed, together with an application, which shows how the Kempf–Ness theorem can be used to derive the singular value decomposition of a finite-dimensional complex operator. A recent generalization of the Kempf–Ness theorem is given, which is due to Azad and Loeb [4]. It is this theorem that enables us to develop a unified approach to balanced realizations, for both the balanced realizations introduced by Moore, as well as the balanced realizations derived from the Kempf–Ness theorem. We then apply the preceding theory in § 3 to prove our main technical result (Theorem 3.2) on the existence and uniqueness properties of norm-minimizing realizations of an arbitrary transfer function $G(s)$. Moore’s balanced realizations for asymptotically stable systems are derived as a special case. The simplest norm to which our main Theorem 3.2 applies is the Euclidean norm, introduced in § 4. Realizations that minimize the Euclidean norm are characterized for arbitrary transfer functions, as well as for symmetric and Hamiltonian transfer functions.

There are some antecedents of our results and techniques in the control theoretic literature. We mention, in particular, the work of Mullis and Roberts [17], [18], Hwang [10], Williamson [27], and Verriest and Gray [24], where (Moore) balanced realizations for asymptotically stable systems are also treated as a minimization problem for a suitable performance measure. However, the techniques used by these authors are quite different than ours. Byrnes and Willems [6] use moment maps in a similar way as is done here to study a least squares estimation problem.

2. The Kempf–Ness theorem. The purpose of this section is to recall an important recent result from invariant theory, which is due to Kempf and Ness [12]. A recent generalization using several complex variable theory has been obtained by Azad and Loeb [4] and plays a central role in our approach to the balanced realization problem. For references on invariant theory, we refer to Dieudonné and Carrell [7], Kraft [14], and Mumford and Fogarty [19].

Recall that a linear algebraic group is called *reductive* if the radical, i.e., the maximal connected solvable normal subgroup, is a torus.

We consider an arbitrary complex reductive algebraic Lie group G with maximal compact subgroup K . Examples of such groups are

- (i) The general linear group $GL(n, \mathbb{C})$ of invertible complex $n \times n$ matrices with maximal compact subgroup $U(n, \mathbb{C})$, the subgroup of $n \times n$ unitary matrices;
- (ii) The special linear group $SL(n, \mathbb{C})$ of invertible complex $n \times n$ matrices with determinant 1. A maximal compact subgroup is $SU(n, \mathbb{C})$, defined by all unitary $n \times n$ matrices of determinant one;
- (iii) The complex orthogonal group $O(n, \mathbb{C})$ of all complex $n \times n$ matrices T satisfying $T \cdot T' = I_n$. A maximal compact subgroup is the orthogonal group $O(n, \mathbb{R})$ of real orthogonal $n \times n$ matrices.

Let

$$(2.1) \quad \begin{aligned} \alpha : G \times V &\rightarrow V, \\ (g, x) &\mapsto g \cdot x \end{aligned}$$

denote a linear algebraic action of G on a finite-dimensional complex vector space V .

Thus α is an algebraic map such that, for all $x \in V$ and $g, h \in G$,

$$(2.2a) \quad e \cdot x = x,$$

$$(2.2b) \quad (gh) \cdot x = g \cdot (h \cdot x),$$

$$(2.2c) \quad \text{the map } V \rightarrow V, x \mapsto g \cdot x, \text{ is } \mathbb{C}\text{-linear}$$

hold (where $e \in G$ denotes the identity element). Given an element $x \in V$, the subset of V

$$(2.3) \quad G \cdot x = \{g \cdot x \mid g \in G\}$$

is called an orbit of G . Since G is a complex manifold, each orbit $G \cdot x$ is a complex manifold that is biholomorphic to the complex homogeneous space G/H , where $H = \{g \in G \mid g \cdot x = x\}$ is the stabilizer group.

We are interested in the critical points of K -invariant smooth functions, defined on G -orbits of an algebraic group action α . Here a function $\varphi : G \cdot x \rightarrow \mathbb{R}$ on a G -orbit $G \cdot x$ is called *K-invariant* if, for all $g \in G$ and all $k \in K$,

$$(2.4) \quad \varphi(kg \cdot x) = \varphi(g \cdot x)$$

holds. A specific example of such a situation arises as follows.

A Hermitian inner product $\langle \cdot, \cdot \rangle$ on V is called *K-invariant* if $\langle k \cdot x, k \cdot y \rangle = \langle x, y \rangle$ holds for all $x, y \in V$ and $k \in K$. This induces a K -invariant Hermitian norm on V , defined by $\|x\|^2 := \langle x, x \rangle$. Fix any such K -invariant Hermitian norm on V . For any given $x \in V$, we consider the following (induced) *distance functions* on $G \cdot x$ and G :

$$(2.5a) \quad \phi_x : G \cdot x \rightarrow \mathbb{R}, \quad g \cdot x \mapsto \|g \cdot x\|^2$$

and

$$(2.5b) \quad \phi_x : G \rightarrow \mathbb{R}, \quad g \mapsto \|g \cdot x\|^2.$$

Note that $\phi_x(g)$ is the (square of the) distance of the transformed vector $g \cdot x$ to the origin.

Let

$$(2.6) \quad \mu(x) := D\phi_x(e)$$

denote the Fréchet derivative of ϕ_x at the identity element e of G . This defines a map

$$(2.7) \quad \mu : V \rightarrow \mathfrak{g}^*, \quad x \mapsto \mu(x)$$

from V to the dual of the Lie algebra \mathfrak{g} of G . Note that μ is K -equivariant with respect to the coadjoint action of K on \mathfrak{g}^* . μ is called the *moment map* for the action (2.1). Thus the zeros of the moment map (2.7) are precisely the critical points of the smooth function defined on the G -orbits $G \cdot x = \{g \cdot x \mid g \in G\}$, which gives the (square of the) distance of an element $y = g \cdot x$ to the origin.

We can now state the Kempf–Ness result. For a proof, we refer to [12], [19]. See [23] for a real version of the theorem.

THEOREM 2.1 (Kempf–Ness). *Let $\alpha : G \times V \rightarrow V$ be an algebraic action of a complex reductive group G on a finite-dimensional vectorspace V and let $\|\cdot\|$ be a K -invariant*

Hermitian norm on V . Then

- (i) *The induced norm function $\phi_x: G \rightarrow \mathbb{R}$ has a global minimum (i.e., $\|g_0 \cdot x\| = \inf_{g \in G} \|g \cdot x\|$ for some $g_0 \in G$) if and only if there exists a critical point of $\phi_x: G \rightarrow \mathbb{R}$, if and only if the G -orbit $G \cdot x$ is a closed subset of V ;*
- (ii) *Let $G \cdot x$ be a closed subset of V . Every critical point of ϕ_x is a global minimum, and the set of global minima of $\phi_x: G \cdot x \rightarrow \mathbb{R}$ is a single, uniquely determined K -orbit.*

The above result has been generalized by Azad and Loeb [4] for plurisubharmonic functions defined on complex homogeneous spaces. To state their result, we recall some basic facts and definitions from several complex variable theory concerning plurisubharmonic functions. A basic reference is [15].

Assume that $X \subset \mathbb{C}^n$ is an open and connected subset of \mathbb{C}^n . An upper semicontinuous function $f: X \rightarrow \mathbb{R} \cup \{-\infty\}$ is called *plurisubharmonic* (*plush*) if the restriction of f to any one-dimensional complex disc is subharmonic, i.e., if, for all $a, b \in \mathbb{C}^n$ and $z \in \mathbb{C}$ with $a + bz \in X$, the function

$$(2.8) \quad z \mapsto f(a + bz)$$

is subharmonic. The class of plurisubharmonic functions constitutes a natural extension of the class of convex functions: Any convex function on X is plurisubharmonic. We list a number of further properties of plurisubharmonic functions.

Properties. (i) Let $f: X \rightarrow \mathbb{C}$ be holomorphic. Then the functions $\log |f|$ and $|f|^p$, $p > 0$ real, are plurisubharmonic (plush).

(ii) Let $f: X \rightarrow \mathbb{R}$ be twice continuously differentiable. Then f is plush if and only if the *Levi form* of f

$$(2.9) \quad L(f) := \left(\frac{\partial^2 f}{\partial z_i \partial \bar{z}_j} \right)$$

is positive semidefinite on X . We say that $f: X \rightarrow \mathbb{R}$ is *strictly plush* if the Levi form $L(f)$ is positive definite, i.e., $L(f) > 0$, on X .

(iii) Let $f, g: X \rightarrow \mathbb{R}$ be plush, $a \geq 0$ real. Then the functions $f + g$ and $a \cdot f$ are plush.

(iv) Let $\varphi: X \rightarrow Y$ be holomorphic and $f: Y \rightarrow \mathbb{R}$ be plush. Then $f \circ \varphi: X \rightarrow \mathbb{R}$ is plush. By property (iv), the notion of plurisubharmonic functions can be extended to functions on any complex manifold (and even on any complex analytic subvariety of \mathbb{C}^n). We then have the next important property;

(v) Let $M \subset X$ be a complex submanifold and $f: X \rightarrow \mathbb{R}$ plush. Then the restriction $f|_M: M \rightarrow \mathbb{R}$ of f to M is plush. Any norm on \mathbb{C}^n is certainly a convex function of its arguments and, therefore, plush. More generally, we have by property (v) the final property;

(vi) Let $\|\cdot\|$ be any norm on \mathbb{C}^n and let X be a complex analytic subvariety of \mathbb{C}^n . Then, for every $a \in \mathbb{C}^n$, the distance function $\phi_a: X \rightarrow \mathbb{R}$, $\phi_a(x) = \|x - a\|$, is plurisubharmonic.

If $\|\cdot\|$ is the induced norm of an Hermitian inner product on \mathbb{C}^n and $X \subset \mathbb{C}^n$ is a complex analytic submanifold, then the distance functions $\phi_a: X \rightarrow \mathbb{R}$, $\phi_a(x) = \|x - a\|^2$, are strictly plurisubharmonic.

For a proof of the following result, we refer to [4].

THEOREM 2.2 (Azad-Loeb). *Let $\alpha: G \times V \rightarrow V$ be an algebraic action of a complex reductive group G on a finite-dimensional complex vectorspace V and let $\varphi: G \cdot x \rightarrow \mathbb{R}$ be a smooth unitarily invariant strictly plurisubharmonic function defined on a G -orbit $G \cdot x$. Suppose that a global minimum of φ exists on $G \cdot x$. Then every critical point of φ is a point where φ assumes its global minimum. The set of global minima is a single K -orbit.*

Since any norm function induced by a K -invariant Hermitian inner product on V is strictly plurisubharmonic on any G -orbit, part (ii) of the Kempf–Ness theorem follows immediately from the Azad–Loeb result.

To see how the above theorems can be applied in a concrete situation, we discuss the following example, which is closely related to the SVD of a finite-dimensional operator.

Example (Singular value decomposition (SVD)). Here $G = GL(n, \mathbb{C})$ and $V = \mathbb{C}^{M \times n} \times \mathbb{C}^{n \times N}$, $n \leq \min(M, N)$. $K = U_n(\mathbb{C})$ and a K -invariant Hermitian norm on V is $(A^* := \bar{A}')$

$$(2.10) \quad \|(X, Y)\|^2 := \text{tr}(X^*X + YY^*),$$

i.e., is given by the sum of the norm squares of the entries of X and Y . The action on V is

$$\begin{aligned} \alpha: GL(n, \mathbb{C}) \times (\mathbb{C}^{M \times n} \times \mathbb{C}^{n \times N}) &\rightarrow \mathbb{C}^{M \times n} \times \mathbb{C}^{n \times N}, \\ (S, (X, Y)) &\mapsto (XS^{-1}, SY). \end{aligned}$$

The induced distance function on $GL(n, \mathbb{C})$ is

$$\phi_{(X, Y)}: GL(n, \mathbb{C}) \rightarrow \mathbb{R}$$

with

$$(2.11) \quad \phi_{(X, Y)}(S) = \text{tr}((S^*)^{-1}X^*XS^{-1}) + \text{tr}(SY Y^*S^*),$$

from which the moment map is easily computed.

We find that

$$(2.12) \quad \mu(X, Y) = 2(YY^* - X^*X),$$

and thus the critical points are characterized by

$$(2.13) \quad \mu(X, Y) = 0 \Leftrightarrow X^*X = YY^*.$$

Let $GL(n, \mathbb{C}) \cdot (X, Y) = \{(XS^{-1}, SY) | S \in GL(n, \mathbb{C})\}$ be an orbit of the $GL(n, \mathbb{C})$ -action α . It is not difficult to show that $GL(n, \mathbb{C}) \cdot (X, Y)$ is a closed subset of $\mathbb{C}^{M \times n} \times \mathbb{C}^{n \times N}$ if and only if $\text{rk}X = \text{rk}Y = \text{rk}XY$; see Kraft [14].

Thus the Kempf–Ness theorem gives the following result.

COROLLARY 2.3. (i) *There exists $S_0 \in GL(n, \mathbb{C})$, which minimizes the distance function $\phi_{(X, Y)}$ defined by (2.11) if and only if $\text{rk}X = \text{rk}Y = \text{rk}XY$;*

(ii) *Let $\text{rk}X = \text{rk}Y = n$. Every critical point of $\phi_{(X, Y)}$ minimizes $\phi_{(X, Y)}$. There exists $S \in GL(n, \mathbb{C})$, unique up to a unitary left factor, such that $(\tilde{X}, \tilde{Y}) := (XS^{-1}, SY)$ satisfies*

$$(2.14) \quad \tilde{X}^* \tilde{X} = \tilde{Y} \tilde{Y}^*.$$

Using the well-known fact that every rank n matrix $A \in \mathbb{C}^{M \times N}$ has a full rank factorization $A = XY$ and that $\{XS^{-1}, SY | S \in GL(n, \mathbb{C})\}$ is the whole class of such factorizations, Corollary 2.3 immediately implies the following result, which is equivalent to the SVD.

COROLLARY 2.4 (SVD). (i) *Every $A \in \mathbb{C}^{M \times N}$ with rank $A = n$ has a factorization $A = XY$, where X and Y are $M \times n$ and $n \times N$ full rank matrices, with*

$$(2.15) \quad X^*X = YY^*;$$

(ii) *If $A = X_1 Y_1 = X_2 Y_2$ are factorizations as in (i), then*

$$X_2 = X_1 T^{-1}, \quad Y_2 = T Y_1$$

for a unique unitary transformation $T \in U(n, \mathbb{C})$.

Remark 2.4. (a) The singular values of an operator A are defined as the (nonnegative) square roots of the eigenvalues of AA^* . Thus the singular values of A coincide with the eigenvalues of the positive definite Hermitian matrix (2.15). It follows that, for $rkX = rkY = rkXY$, the minimal value of the distance function (2.11) is given by the $2 \times$ (sum of the singular values of XY).

(b) Part (i) of Corollary 2.3 is equivalent to a result of Flanders [8]. Flanders's proof, as well as those of Anderson and Olkin [2] and Wimmer [28], is, however, quite different and longer than the above proof.

3. Norm and function balanced realizations. Consider the complex vector space of triples (A, B, C)

$$(3.1) \quad L(n, m, p) := \{(A, B, C) \in \mathbb{C}^{n \times n} \times \mathbb{C}^{n \times m} \times \mathbb{C}^{p \times n}\}.$$

The reductive group $GL(n, \mathbb{C})$ of complex invertible $n \times n$ matrices S acts on $L(n, m, p)$ via the algebraic group action

$$(3.2) \quad \begin{aligned} \sigma: GL(n, \mathbb{C}) \times L(n, m, p) &\rightarrow L(n, m, p), \\ (S, (A, B, C)) &\mapsto (SAS^{-1}, SB, CS^{-1}). \end{aligned}$$

The orbits of σ

$$(3.3) \quad \mathcal{O}(A, B, C) = \{(SAS^{-1}, SB, CS^{-1}) \mid S \in GL(n, \mathbb{C})\}$$

are complex homogeneous spaces and thus complex submanifolds of $L(n, m, p)$.

A function $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}$ is called *unitarily invariant* if, for all unitary matrices $S \in U(n, \mathbb{C})$, $SS^* = I_n$,

$$(3.4) \quad f(SAS^{-1}, SB, CS^{-1}) = f(A, B, C)$$

holds. We are interested in the critical point structure of smooth, unitarily invariant plurisubharmonic functions $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}$ on $GL(n, \mathbb{C})$ -orbits $\mathcal{O}(A, B, C)$. A particular case of interest is where the function $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}$ is induced from a suitable norm on $L(n, m, p)$.

Thus let $\langle \cdot, \cdot \rangle$ denote a Hermitian inner product on the \mathbb{C} -vector space $L(n, m, p)$. The induced Hermitian norm of (A, B, C) is defined by

$$(3.5) \quad \|(A, B, C)\|^2 = \langle (A, B, C), (A, B, C) \rangle.$$

A Hermitian norm (3.5) is called unitarily invariant if

$$(3.6) \quad \|(SAS^{-1}, SB, CS^{-1})\| = \|(A, B, C)\|$$

holds for all unitary transformations S , $SS^* = I_n$, and $(A, B, C) \in L(n, m, p)$. Any Hermitian norm (3.5) defines a smooth strictly plurisubharmonic function

$$(3.7) \quad \begin{aligned} \phi: \mathcal{O}(A, B, C) &\rightarrow \mathbb{R}, \\ (SAS^{-1}, SB, CS^{-1}) &\mapsto \|(SAS^{-1}, SB, CS^{-1})\|^2 \end{aligned}$$

on $\mathcal{O}(A, B, C)$.

In the following, we fix a strictly proper transfer function $G(s) \in \mathbb{C}^{p \times m}(s)$ of McMillan degree n , and we also fix an initial controllable and observable realization $(A, B, C) \in L(n, m, p)$ of $G(s)$ as follows:

$$(3.8) \quad G(s) = C(sI - A)^{-1}B.$$

Thus the $GL(n, \mathbb{C})$ -orbit $\mathcal{O}(A, B, C)$ parametrizes the set of all (minimal) realizations of $G(s)$.

Our goal is to study the variation of the norm $\|(SAS^{-1}, SB, CS^{-1})\|^2$ as S varies in $GL(n, \mathbb{C})$. In particular, we seek to obtain answers to the following questions:

1. Given a function $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}$, does there exist a realization of $G(s)$ that minimizes f ?

2. How can we characterize the set of realizations of a transfer function that minimize $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}$?

As we will see, the Kempf–Ness and Azad–Loeb theorems give a rather general solution to these questions. Let $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}$ be a smooth function on $\mathcal{O}(A, B, C)$ and let $\|\cdot\|$ denote a Hermitian norm defined on $L(n, m, p)$.

DEFINITION 3.1. A realization

$$(3.9a) \quad (F, G, H) = (S_0AS_0^{-1}, S_0B, CS_0^{-1})$$

of a transfer function $G(s) = C(sI - A)^{-1}B$ is called *norm balanced* and *function balanced*, respectively, if the function

$$(3.9b) \quad \begin{aligned} \phi: GL(n, \mathbb{C}) &\rightarrow \mathbb{R}, \\ S &\mapsto \|(SAS^{-1}, SB, CS^{-1})\|^2 \end{aligned}$$

or the function

$$(3.10) \quad \begin{aligned} \mathcal{F}: GL(n, \mathbb{C}) &\rightarrow \mathbb{R}, \\ S &\mapsto f(SAS^{-1}, SB, CS^{-1}), \end{aligned}$$

respectively, has a critical point at $S = S_0$; i.e., if the Fréchet derivative

$$(3.11) \quad D\phi|_{S_0} = 0,$$

respectively,

$$(3.12) \quad D\mathcal{F}|_{S_0} = 0$$

vanishes. (F, G, H) is called *norm minimal* or *function minimizing* if $\phi(S_0)$, respectively, $\mathcal{F}(S_0)$, is a global minimum for the function (3.9) or (3.10) on $GL(n, \mathbb{C})$. See Fig. 1.

We need the following characterization of controllable and observable realizations as the $GL(n, \mathbb{C})$ -stable points for the similarity action $(A, B, C) \mapsto (SAS^{-1}, SB, CS^{-1})$.

LEMMA 3.2. $(A, B, C) \in L(n, m, p)$ is controllable and observable if and only if the following conditions are satisfied:

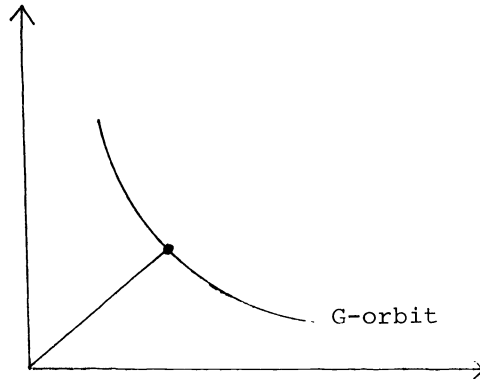


FIG. 1. Norm minimality.

(i) *The similarity orbit $\mathcal{O}(A, B, C) := \{(SAS^{-1}, SB, CS^{-1}) | S \in GL(n, \mathbb{C})\}$ is a closed subset of $L(n, m, p)$;*

(ii) $\dim_{\mathbb{C}} \mathcal{O}(A, B, C) = n^2$.

Proof. The necessity of (ii) is obvious, since $GL(n, \mathbb{C})$ acts freely on controllable and observable triples via similarity. The necessity of (i) follows from realization theory, since $\mathcal{O}(A, B, C)$ is a fibre of the continuous map

$$(3.13) \quad \begin{aligned} \mathcal{H}: L(n, m, p) &\rightarrow \prod_{i=1}^{\infty} \mathbb{C}^{p \times m}, \\ (F, G, H) &\mapsto (HF^i G | i \in \mathbb{N}_0), \end{aligned}$$

where $\prod_{i=1}^{\infty} \mathbb{C}^{p \times m}$ is endowed with the product topology.

To prove the sufficiency, let us assume that, e.g., (A, B) is not controllable while conditions (i), (ii) are satisfied. Without loss of generality,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad C = [C_1, C_2],$$

and (A_{11}, B_1) controllable, $A_{ii}n_i \times n_i$ for $i = 1, 2$. Consider the one-parameter group of transformations

$$S_t := \begin{bmatrix} I_{n_1} & 0 \\ 0 & t^{-1}I_{n_2} \end{bmatrix} \in GL(n, \mathbb{C})$$

for $t \neq 0$. Then $(A_t, B_t, C_t) := (S_t A S_t^{-1}, S_t B, C S_t^{-1}) \in \mathcal{O}(A, B, C)$ with

$$A_t = \begin{bmatrix} A_{11} & tA_{12} \\ 0 & A_{22} \end{bmatrix}, \quad B_t := \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad C_t := [C_1, tC_2].$$

Since $\mathcal{O}(A, B, C)$ is closed,

$$(A_0, B_0, C_0) = \left(\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, [C_1, 0] \right) \in \mathcal{O}(A, B, C),$$

which is stabilized by $S_t, t \in \mathbb{C}^*$. This is a contradiction to (ii). Thus (A, B) must be controllable, and, similarly, (A, C) must be observable. This proves Lemma 3.2. \square

The following theorems are the main results of this paper. They are immediate consequences of Lemma 3.2 and Theorem 2.2 (Azad-Loeb) and Theorem 2.1 (Kempf-Ness), respectively. Recall that a continuous function $f: X \rightarrow \mathbb{R}$ on a topological space X is called *proper* if the inverse image $f^{-1}([a, b])$ of any compact interval $[a, b] \subset \mathbb{R}$ is a compact subset of X . For every proper map $f: X \rightarrow \mathbb{R}$, the image $f(X)$ is a closed subset of \mathbb{R} .

THEOREM 3.3. *Let $G(s) = C(sI - A)^{-1}B \in \mathbb{C}^{p \times m}(s)$ be a strictly proper rational transfer function of McMillan degree n . Let $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}_+ (\mathbb{R}_+ = [0, \infty[)$ be a smooth, unitarily invariant, strictly plurisubharmonic function on $\mathcal{O}(A, B, C)$, which is proper. Then*

- (a) *There exists a global minimum of f in $\mathcal{O}(A, B, C)$;*
- (b) *A controllable and observable realization (A, B, C) of $G(s)$ is function minimal if and only if it is function balanced;*
- (c) *If $(A_1, B_1, C_1), (A_2, B_2, C_2) \in \mathcal{O}(A, B, C)$ are minima of f , then there exists a uniquely determined unitary transformation $S \in U(n, \mathbb{C})$ such that*

$$(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1}).$$

THEOREM 3.4. Let $\|\cdot\|^2$ be a unitarily invariant Hermitian norm on $L(n, m, p)$ and let $G(s) \in \mathbb{C}^{p \times m}(s)$ denote a strictly proper rational transfer function of McMillan degree n . Then

- (a) There exists a norm minimal realization (A, B, C) of $G(s)$;
- (b) A controllable and observable realization $(A, B, C) \in L(n, m, p)$ of $G(s)$ is norm minimal if and only if it is norm balanced;
- (c) If $(A_1, B_1, C_1), (A_2, B_2, C_2) \in L(n, m, p)$ are controllable and observable norm minimal realizations of $G(s)$, then there exists a uniquely determined unitary transformation $S \in U(n, \mathbb{C})$ such that

$$(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1}).$$

Remark 3.5. (1) Let $G(s) \in \mathbb{R}^{p \times m}(s)$ denote a strictly proper *real* rational transfer function of McMillan degree n and let $\|\cdot\|^2$ be a unitarily invariant Hermitian norm on the complex vector space $L(n, m, p)$ (or let $f: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}_+$ be a smooth, unitarily invariant, strictly plurisubharmonic function on the complex similarity orbit that is proper and invariant under complex conjugation: $f(\bar{F}, \bar{G}, \bar{H}) = f(F, G, H)$ for all $(F, G, H) \in \mathcal{O}(A, B, C)$). Then Theorems 3.3 and 3.4 remain valid if conditions (a)–(c) in Theorems 3.3 and 3.4 are replaced by their respective versions (a')–(c') as follows (when we use “norm minimal,” we are referring to Theorem 3.4; “function minimal” refers to Theorem 3.3):

- (a') A *real* controllable and observable realization $(A, B, C) \in L(n, m, p)$ of $G(s)$ is norm (function) minimal if and only if it is norm (function) balanced;
- (b') There exists a *real* norm (function) minimal realization (A, B, C) of $G(s)$;
- (c') If $(A_1, B_1, C_1), (A_2, B_2, C_2) \in L(n, m, p)$ are *real* controllable and observable norm (function) minimal realizations of $G(s)$, then there exists a uniquely determined *real orthogonal* transformation $S \in O(n, \mathbb{R})$ such that

$$(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1}).$$

This follows from the easily established fact that any similarity transformation $S \in GL(n, \mathbb{C})$, which transforms a given real controllable and observable realization (A_1, B_1, C_1) into a real realization $(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1})$, is necessarily real, i.e., $S \in GL(n, \mathbb{R})$. Alternatively, the result follows immediately from Slodowy's real version of Theorem 2.1 [23].

(2) The norm balanced realizations were also considered by Verriest [26], where they are referred to as “optimally clustered.” Also, Verriest points out the invariance of the norm under orthogonal transformations but does not show global minimality of the norm balanced realizations. An example is given that shows not every similarity orbit $\mathcal{O}(A, B, C)$ allows a norm balanced realization. In fact, by Theorem 2.1 (Kempf–Ness) (i), there exists a norm balanced realization in $\mathcal{O}(A, B, C)$ if and only if $\mathcal{O}(A, B, C)$ is a closed subset of $L(n, m, p)$.

Balanced realizations for the class of asymptotically stable linear systems were first introduced by Moore [16] and are defined by the condition that the controllability and observability gramians are equal and diagonal. We will now show that these balanced realizations can be treated as a special case of our above theorems. For simplicity, we consider only the discrete-time case and complex systems (A, B, C) .

DEFINITION 3.5. A complex realization (A, B, C) is called *N-balanced* if and only if

$$(3.14) \quad \sum_{k=0}^N A^k B B^* (A^*)^k = \sum_{k=0}^N (A^*)^k C^* C A^k.$$

An asymptotically stable realization (A, B, C) (i.e., $\sigma(A) < 1$) is said to be *balanced*, or ∞ -balanced, if and only if

$$(3.15) \quad \sum_{k=0}^{\infty} A^k B B^* (A^*)^k = \sum_{k=0}^{\infty} (A^*)^k C^* C A^k.$$

Note that the above terminology differs slightly from the usual terminology in the sense that the controllability, respectively, observability, gramians are not required to be diagonal. Of course, this can always be achieved by an orthogonal change of basis in the state space. In Verriest and Gray [24], realizations (A, B, C) satisfying (3.14) or (3.15) are called essentially balanced.

To prove the existence of (N) -balanced realizations, we consider the minimization problem for the “ N -gramian norm” function

$$(3.16) \quad f_N(A, B, C) := \text{tr} \sum_{k=0}^N (A^k B B^* (A^*)^k + (A^*)^k C^* C A^k)$$

for $N \in \mathbb{N} \cup \{\infty\}$. Let $(A, B, C) \in L(n, m, p)$ be a controllable and observable realization with $N \geq n$. Consider the smooth unitarily invariant function on the $GL(n, \mathbb{C})$ -orbit

$$(3.17) \quad \begin{aligned} f_N: \mathcal{O}(A, B, C) &\rightarrow \mathbb{R}_+, \\ (SAS^{-1}, SB, CS^{-1}) &\mapsto f_N(SAS^{-1}, SB, CS^{-1}), \end{aligned}$$

where $f_N(SAS^{-1}, SB, CS^{-1})$ is defined by (3.16) for any $N \in \mathbb{N} \cup \{\infty\}$. It is not difficult to show that $f_N: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}_+$ is proper for $N \geq n$; see Perkins, Helmke, and Moore [22]. Using properties (i)–(v) for plurisubharmonic functions that follow Theorem 2.1, it is not difficult to show that f_N is strictly plurisubharmonic for any $N \in \mathbb{N} \cup \{\infty\}$, $N \geq n$.

Thus we can apply Theorem 3.3. To compute the critical points of $f_N: \mathcal{O}(A, B, C) \rightarrow \mathbb{R}_+$, we consider the induced function on $GL(n, \mathbb{C})$

$$(3.18) \quad \begin{aligned} F_N: GL(n, \mathbb{C}) &\rightarrow \mathbb{R}_+, \\ S &\mapsto F_N(SAS^{-1}, SB, CS^{-1}) \end{aligned}$$

for any $N \in \mathbb{N} \cup \{\infty\}$. A simple calculation of the gradient vector ∇F_N at $S = I_n$ shows that

$$(3.19) \quad \nabla F_N(I_n) = 2 \sum_{k=0}^N (A^k B B^* (A^*)^k - (A^*)^k C^* C A^k)$$

for any $N \in \mathbb{N} \cup \{\infty\}$. We conclude the following result.

COROLLARY 3.6. *Given a complex rational strictly proper transfer function $G(s)$ of McMillan degree n , then, for all finite $N \geq n$, there exists a realization (A_*, B_*, C_*) of $G(s)$ that is N -balanced. If (A_1, B_1, C_1) , (A_2, B_2, C_2) are N -balanced realizations of $G(s)$ of order n , $N \geq n$, then*

$$(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1})$$

for a uniquely determined unitary transformation $S \in U(n, \mathbb{C})$. (A_1, B_1, C_1) is N -balanced if and only if it minimizes the N -gramian norm taken over all realizations of $G(s)$ of order n .

This also follows immediately from Corollary 2.3 and from the SVD (Corollary 2.4), applied to the factorization of the $N \times N$ -block Hankel $\mathcal{H}_N = O_N \cdot R_N$, where $R_N := (B, AB, \dots, A^N B)$, $O_N := (C', A'C', \dots, (A')^N C')'$.

For asymptotically stable linear systems, we obtain the following amplification of Moore's fundamental existence and uniqueness theorem for balanced realizations.

THEOREM 3.7. *Given a complex rational strictly proper transfer function $G(s)$ of McMillan degree n and with all poles in the open unit disc, there exists a balanced realization (A_1, B_1, C_1) of $G(s)$ of order n . If $(A_1, B_1, C_1), (A_2, B_2, C_2)$ are two balanced realizations of $G(s)$ of order n , then*

$$(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1})$$

for a uniquely determined unitary transformation $S \in U(n, \mathbb{C})$. An n -dimensional realization (A_1, B_1, C_1) of $G(s)$ is balanced if and only if (A_1, B_1, C_1) minimizes the gramian norm (3.17) (for $N = \infty$), taken over all realizations of $G(s)$ of order n .

We consider a (discrete time) asymptotically stable transfer function $G(s) \in \mathbb{C}^{p \times m}(s)$ of McMillan degree n . Let $\sigma_1 \geq \dots \geq \sigma_n$ denote the singular values of the induced Hankel operator; see Moore [16], Glover [9]. For any $p \in \mathbb{N}$, let

$$(3.20) \quad S_p(G) := \left(\sum_{j=1}^n \sigma_j^p \right)^{1/p}$$

and, for any stable (A, B, C) , let

$$|(A, B, C)|_p := (\text{tr}(W_C^p + W_O^p))^{1/p},$$

where

$$W_C := \sum_{k=0}^{\infty} A^k B B^* (A^*)^k,$$

respectively,

$$W_O := \sum_{k=0}^{\infty} (A^*)^k C^* C A^k,$$

are the controllability, respectively, observability, gramians. The following result has been shown by Williamson [27], in the special case where $p = 1$. His proof is quite different than ours and is based on an inequality of Mullis and Roberts [17].

THEOREM 3.8. *Let (A, B, C) be a controllable and observable realization of the asymptotically stable transfer function $G(s)$. Then, for all $p \in \mathbb{N}$,*

$$|(A, B, C)|_p \geq 2^{1/p} \cdot S_p(G).$$

(A, B, C) is balanced if and only if, for one $p \in \mathbb{N}$ (and hence for all p),

$$(3.21) \quad |(A, B, C)|_p = 2^{1/p} \cdot S_p(G),$$

where $S_p(G)$ is defined by (3.20).

Proof. A straightforward computation of the gradient vector of the function $\phi: GL(n, \mathbb{C}) \rightarrow \mathbb{R}, S \mapsto |(SAS^{-1}, SB, CS^{-1})|_p^p$ at $S = I_n$ gives

$$(3.22) \quad \nabla \Phi(I_n) = 2p(W_C^p - W_O^p).$$

Thus (A, B, C) is function balanced for $||_p$ if and only if $W_C(A, B, C)^p = W_O(A, B, C)^p$, i.e., if and only if (A, B, C) is balanced. Thus the minimal value of $|(A, B, C)|_p$ is achieved (where (A, B, C) runs through all controllable and observable realizations of $G(s)$) if and only if (A, B, C) is balanced (this uses Theorem 3.7). In the balanced case,

$$|(A, B, C)|_p = 2^{1/p} \cdot (\text{tr } W_C^p)^{1/p}.$$

By Remark 2.4, $\text{tr } W_C = S_1(G)$ and, similarly, $\text{tr } W_C^p = [S_p(G)]^p$. The result follows. \square

4. New classes of balanced realizations. The simplest candidate of a unitarily invariant Hermitian norm on $L(n, m, p)$ is the *standard Euclidean norm*, defined by

$$(4.1) \quad \|(A, B, C)\|^2 := \operatorname{tr} AA^* + \operatorname{tr} BB^* + \operatorname{tr} C^*C,$$

where $X^* = \bar{X}'$ denotes Hermitian transpose. An application of Theorem 3.4 to this norm yields the following result, which describes a new class of norm minimal realizations.

THEOREM 4.1. *Let $G(s)$ be a real rational strictly proper transfer function of McMillan degree n . Then*

- (i) *There exists a (real) controllable and observable realization (A, B, C) of $G(s)$ with*

$$(*) \quad AA' + BB' = A'A + C'C.$$

- (ii) *If $(A_1, B_1, C_1), (A_2, B_2, C_2)$ are (real) controllable and observable realizations of $G(s)$ satisfying $(*)$, then there exists a unique orthogonal transformation $S \in O(n, \mathbb{R})$, with*

$$(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1}).$$

- (iii) *An n -dimensional realization (A, B, C) of $G(s)$ satisfies $(*)$ if and only if it minimizes the Euclidean norm (4.1), taken over all possible n -dimensional realizations of $G(s)$.*

Proof. For any controllable and observable realization (A, B, C) of $G(s)$, consider the function $\phi: GL(n, \mathbb{R}) \rightarrow \mathbb{R}$ defined by the Euclidean norm

$$\phi(S) = \|(SAS^{-1}, SB, CS^{-1})\|^2.$$

Thus $\phi(S) = \operatorname{tr}(SAS^{-1}(S')^{-1}A'S') + \operatorname{tr}(SBB'S') + \operatorname{tr}((S')^{-1}C'CS^{-1})$. The gradient vector of ϕ at $S = I_n$ is

$$(4.2) \quad \nabla \phi(I_n) = 2(AA' - A'A + BB' - C'C),$$

and thus (A, B, C) is norm balanced for the Euclidean norm (4.1) if and only if

$$AA' - A'A + BB' - C'C = 0,$$

which is equivalent to $(*)$. The result now follows immediately from Theorem 3.2. \square

Similar results hold for symmetric or Hamiltonian transfer functions. Recall that a real rational $m \times m$ -transfer function $G(s)$ is called *symmetric*, respectively, *Hamiltonian*, if for all $s \in \mathbb{C}$

$$(4.3) \quad G(s) = G(s)',$$

respectively,

$$(4.4) \quad G(s) = G(-s)'$$

(where $'$ denotes transpose).

Every strictly proper symmetric transfer function $G(s)$ of McMillan degree n has a minimal *signature symmetric* realization (A, B, C) satisfying

$$(4.5) \quad (AI_{pq})' = AI_{pq}, \quad C' = I_{pq}B,$$

where $I_{pq} = \operatorname{diag}(\varepsilon_1, \dots, \varepsilon_n)$ with

$$(4.6) \quad \varepsilon_i = \begin{cases} 1 & i = 1, \dots, p, \\ -1 & i = p+1, \dots, p+q = n. \end{cases}$$

Here $p - q$ is the Cauchy–Maslov index of $G(s)$; cf. Anderson and Bitmead [1], Byrnes and Duncan [5]. Similarly, every strictly proper Hamiltonian transfer function $G(s) \in \mathbb{R}(s)^{m \times m}$ of McMillan degree $2n$ has a minimal Hamiltonian realization (A, B, C) satisfying

$$(4.7) \quad (AJ)' = AJ, \quad C' = JB,$$

where

$$(4.8) \quad J = \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}$$

is the standard complex structure. Let $O(p, q)$, respectively, $Sp(n, \mathbb{R})$, denote the (real) isotropy groups of I_{pq} , respectively, J , i.e.,

$$(4.9a) \quad T \in O(p, q) \Leftrightarrow T' I_{pq} T = I_{pq}, \quad T \in GL(n, \mathbb{R}),$$

$$(4.9b) \quad T \in Sp(n, \mathbb{R}) \Leftrightarrow T' J T = J, \quad T \in GL(2n, \mathbb{R}).$$

THEOREM 4.2. (i) *Every strictly proper symmetric transfer function $G(s) \in \mathbb{R}(s)^{m \times m}$ with McMillan degree n and Cauchy–Maslov index $p - q$ has a controllable and observable signature symmetric realization (A, B, C) satisfying*

$$(4.10a) \quad (A I_{pq})' = A I_{pq}, \quad C' = I_{pq} B,$$

$$(4.10b) \quad A A' + B B' = A' A + C' C.$$

(ii) *If $(A_1, B_1, C_1), (A_2, B_2, C_2)$ are two minimal realizations of $G(s)$ satisfying (4.10a) and (4.10b), then there exists a unique orthogonal transformation $S = \text{diag}(S_1, S_2) \in O(p) \times O(q) \subset O(n)$ with*

$$(A_2, B_2, C_2) = (S A_1 S^{-1}, S B_1, C_1 S^{-1}).$$

A similar result holds for Hamiltonian transfer functions.

THEOREM 4.3. (i) *Every strictly proper Hamiltonian transfer function $G(s) \in \mathbb{R}(s)^{m \times m}$ with McMillan degree $2n$ has a controllable and observable Hamiltonian realization (A, B, C) satisfying*

$$(4.11a) \quad (AJ)' = AJ, \quad C' = JB,$$

$$(4.11b) \quad A A' + B B' = A' A + C' C.$$

(ii) *If $(A_1, B_1, C_1), (A_2, B_2, C_2)$ are two minimal realizations of $G(s)$ satisfying (4.11a) and (4.11b), then there exists a unique symplectic transformation $S \in Sp(n, \mathbb{R}) \cap O(2n; \mathbb{R})$ with*

$$(A_2, B_2, C_2) = (S A_1 S^{-1}, S B_1, C S^{-1}).$$

Proofs. We first prove Theorem 4.3. Let $Sp(n, \mathbb{C})$ denote the complex symplectic group; i.e., $T \in Sp(n, \mathbb{C})$ if and only if $T \in GL(2n, \mathbb{C})$ and $T' J T = J$. $Sp(n, \mathbb{C})$ is a reductive Lie group with maximal compact subgroup $K = Sp(n; \mathbb{C}) \cap U_{2n}(\mathbb{C})$. Let V denote the complex vector space of all triples $(A, B, C) \in \mathbb{C}^{2n \times 2n} \times \mathbb{C}^{2n \times m} \times \mathbb{C}^{m \times 2n}$ satisfying $(AJ)' = AJ, JB = C'$. The Euclidean norm (4.1) defines a K -invariant Hermitian norm on V . By Lemma 3.3, the $Sp(n, \mathbb{C})$ -orbit of a point $(A, B, C) \in V$ is closed if (A, B, C) is controllable and observable. Theorem 4.3 now follows immediately from Theorem 2.2 (Kempf–Ness), once it is observed that Remark 3.4 remains in force (with the appropriate modification) and that $Sp(n, \mathbb{R}) \cap O(2; \mathbb{R})$ is the set of real points of K .

Unfortunately, the above direct proof does not work for Theorem 4.2, since $O(p, q)$ is not reductive. We therefore proceed in a different way. We first prove (i). By Theorem

4.1, there exists a minimal realization (A, B, C) of the symmetric transfer function $G(s)$, which satisfies (4.10b), and (A, B, C) is unique up to an orthogonal change of basis $S \in O(n; \mathbb{R})$. By the symmetry of $G(s)$, with (A, B, C) , (A', C', B') is also a realization of $G(s)$. Note that if (A, B, C) satisfies (4.10b), (A', C', B') also satisfies (4.10b). Thus, by the above uniqueness property (ii) of Theorem 4.1, there exists a unique $S \in O(n; \mathbb{R})$ with

$$(4.12) \quad (A', C', B') = (SAS', SB, CS').$$

By transposing (4.12), we also obtain that

$$(4.13) \quad (A, B, C) = (SA'S', SC', B'S')$$

or, equivalently,

$$(4.14) \quad (A', C', B') = (S'AS, S'B, CS).$$

Thus (by minimality of (A, B, C)) $S = S'$ is symmetric orthogonal. A straightforward argument (see Byrnes and Duncan [5] for details) shows that the signature of S is equal to the Cauchy–Maslov index $p - q$ of $G(s) = C(sI - A)^{-1}B$. Thus $S = T'I_{pq}T$ for $T \in O(n; \mathbb{R})$, and the new realization

$$(F, G, H) := (TAT', TB, CT')$$

satisfies (4.10a), (4.10b).

To prove part (ii) of Theorem 4.2, let

$$(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1})$$

be two realizations of $G(s)$ that satisfy (4.10a), (4.10b). Byrnes and Duncan [5, Thm. 4.1] show that (4.10a) implies that $S \in O(p, q)$. By Theorem 4.1(ii), also $S \in O(n, \mathbb{R})$. Since $O(p) \times O(q) = O(p, q) \cap O(n, \mathbb{R})$, the result follows. \square

Remark 4.4. (1) There is a formal analogy between norm minimal realizations (A, B, C) satisfying

$$(**) \quad AA' - A'A + BB' - C'C = 0$$

and balanced realizations obtained by the Riccati equation; see Jonkheere and Silverman [11]. Riccati balanced realizations are defined by the following result, whose proof is immediate, by letting T denote the positive definite square root of the uniquely determined positive definite stabilizing solution of the algebraic Riccati equation; see [25] and [11] for details.

PROPOSITION. *Given any controllable and observable n -dimensional realization $(\tilde{A}, \tilde{B}, \tilde{C})$, there exists a state space equivalent realization $(A, B, C) = (T\tilde{A}T^{-1}, T\tilde{B}, \tilde{C}T^{-1})$ satisfying*

$$(***) \quad A + A' + BB' - C'C = 0.$$

If (A_i, B_i, C_i) , $i = 1, 2$ are two such realizations, there exists a unique orthogonal transformation $S \in O(n, \mathbb{R})$ with $(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1})$.

Note the apparent similarity between formulas (**) and (***).

(2) Euclidean norm balanced realizations (**) of Theorem 4.1 were also considered by Verriest [26]. See part (2) of Remark 3.5.

Acknowledgments. The author thanks the two anonymous referees for their careful review of this paper.

REFERENCES

- [1] B. D. O. ANDERSON AND R. R. BITMEAD, *The matrix Cauchy index: Properties and applications*, SIAM J. Appl. Math., 33 (1977), pp. 655–672.

- [2] T. W. ANDERSON AND I. OLKIN, *An extremal problem for positive definite matrices*, Linear and Multilinear Algebra, 6 (1978), pp. 257–262.
- [3] V. I. ARNOLD, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, New York, 1989.
- [4] H. AZAD AND J. J. LOEB, *On a theorem of Kempf and Ness*, Indiana Univ. Math. J., 39 (1990), pp. 61–65.
- [5] C. I. BYRNES AND T. E. DUNCAN, *On certain topological invariants arising in system theory*, in New Directions in Applied Mathematics, P. J. Hilton and G. S. Young, eds., Springer-Verlag, Berlin, New York, 1989, pp. 29–72.
- [6] C. I. BYRNES AND J. C. WILLEMS, *Least squares estimation, linear programming and momentum*, IMA J. Math. Control Inform., 3 (1986), pp. 103–118.
- [7] J. DIEUDONNÉ AND J. B. CARRELL, *Invariant theory, old and new*, in Advances in Mathematics, Academic Press, New York, 1971.
- [8] H. FLANDERS, *An extremal problem on the space of positive definite matrices*, Linear and Multilinear Algebra, 3 (1975), pp. 33–39.
- [9] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.
- [10] S. Y. HWANG, *Minimum uncorrelated unit noise in state space digital filtering*, IEEE Trans. Acoust. Speech Signal Process., ASSP-25 (1977), pp. 273–281.
- [11] E. A. JONCKHEERE AND L. M. SILVERMAN, *A new set of invariants for linear systems—application to reduced order compensation design*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 953–964.
- [12] G. KEMPF AND L. NESS, *The length of vectors in representation spaces*, in Algebraic Geometry, K. Lonsted, ed., Lecture Notes in Math., 732, Springer-Verlag, Berlin, New York, 1979, pp. 233–244.
- [13] C. KENNEY AND G. HEWER, *Necessary and sufficient conditions for balancing unstable systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 157–160.
- [14] H. KRAFT, *Geometrische Methoden in der Invariantentheorie*, Aspects of Mathematics, D1, Vieweg, Braunschweig, Germany, 1984.
- [15] S. G. KRANTZ, *Function Theory of Several Complex Variables*, John Wiley, New York, 1982.
- [16] B. C. MOORE, *Principal component analysis in linear systems: Controllability, observability and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.
- [17] C. T. MULLIS AND R. A. ROBERTS, *Synthesis of minimum roundoff noise fixed point digital filters*, IEEE Trans. Circuits Sys., C AS-23 (1976), pp. 551–562.
- [18] ———, *Roundoff noise in digital filters: Frequency transformations and invariants*, IEEE Trans. Acoust. Speech Signal Process., ASSP-24 (1976), pp. 538–550.
- [19] D. MUMFORD AND J. FOGARTY, *Geometric Invariant Theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete 34, 2nd ed., Springer-Verlag, Berlin, Germany, 1982.
- [20] L. NESS, *A stratification of the null cone via the moment map*, Amer. J. Math. (1984), pp. 1281–1325.
- [21] R. J. OBER, *Balanced realizations: Canonical form, parametrization, model reduction*, Internat. J. Control, 46 (1987), pp. 643–670.
- [22] J. E. PERKINS, U. HELMKE AND J. B. MOORE, *Balanced realizations via gradient flow techniques*, Systems Control Lett., 14 (1990), pp. 369–380.
- [23] P. SŁODOWY, *Zur Geometrie der Bahnen reell reduktiver Gruppen*, in Algebraic Transformation Groups and Invariant Theory, H. Kraft, P. Slodowy, and Springer, eds., Birkhäuser, Boston, 1989, pp. 133–144.
- [24] E. I. VERRIEST AND W. S. GRAY, *Robust design problems: A geometric approach*, in Linear Circuits, Systems and Signal Processing: Theory and Applications, Byrnes, Martin, and Saeks, eds., North-Holland, Amsterdam, 1988, pp. 321–328.
- [25] E. I. VERRIEST, *Suboptimal LQG-design via balanced realizations*, in Proc. 20th IEEE Conf. on Decision and Control, 1981, pp. 686–687.
- [26] ———, *Minimum sensitivity implementation for multi-mode systems*, in Proc. IEEE Conf. on Decision and Control, Austin, TX, 1988, pp. 2165–2170.
- [27] D. WILLIAMSON, *A property of internally balanced and low noise structures*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 633–634.
- [28] K. WIMMER, *Extremal problems for Hölder norms of matrices and realizations of linear systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 314–322.

EXISTENCE THEOREMS FOR POSITIVE SEMIDEFINITE AND SIGN INDEFINITE STABILIZING SOLUTIONS OF H_∞ RICCATI EQUATIONS*

GARY HEWER†

Abstract. The existence of both positive semidefinite and sign indefinite stabilizing solutions of H_∞ -type Riccati equations is determined via joint eigenvalue and Hamiltonian tests. These tests determine the inertia properties of the solutions. They are dependent on a formula—derived in this paper—that shows how the solution of game Riccati equations can be decomposed into a sum of two well-known Riccati equations, namely, the standard filter Riccati equation and the bounded real Riccati equation. Other properties of stabilizing solutions such as solution partial order, spectral radius monotonicity, and rank are also discussed.

Key words. H_∞ -optimal control, Riccati equations, bounded real system

AMS(MOS) subject classifications. 93D15, 93C05, 93B25

1. Introduction. This paper presents a collection of results on the game Riccati equation

$$(1.1) \quad A^T X + XA + X \left(\frac{1}{\gamma^2} B_1 B_1^T - B_2 B_2^T \right) X + C^T C = 0.$$

The existence of a stabilizing positive definite or positive semidefinite solution $X(\gamma)$ of the game Riccati equation with real parameter γ defined on some infinite half-line $0 < \gamma \leq \infty$ is a prerequisite for an H_∞ -control problem to have a solution [4]. Here A , B_1 , B_2 , and C are real $n \times n$, $n \times p$, $n \times m$, and $q \times n$ real matrices, respectively, and A^T denotes the matrix transpose. It is the objective of this paper to show that many properties of $X(\gamma)$ can be connected to the standard filter algebraic Riccati equation (FARE)

$$(1.2) \quad AZ + ZA^T - ZC^T CZ + B_2 B_2^T = 0$$

and to the bounded real Riccati equation (BRRE)

$$(1.3) \quad (A - ZC^T C)W + W(A - C^T CZ)^T - WC^T CW - \frac{1}{\gamma^2} B_1 B_1^T = 0$$

by noting that any solution of the “dual” game Riccati equation

$$(1.4) \quad AY + YA^T - YC^T CY + B_2 B_2^T - \frac{1}{\gamma^2} B_1 B_1^T = 0$$

can be written as the sum of the appropriate solution of FARE and BRRE (provided they all exist). The technique of expressing the solution of the dual Riccati equation (1.4) as the sum of two terms, one a solution to a filter Riccati equation and one a solution of a bounded real Riccati equation, can be found within the proofs presented in Willems [25], with a subsequent appearance in Molinari [16].

While real symmetric stabilizing solutions of the control algebraic Riccati equation (CARE) are always positive semidefinite and maximal [13], there can exist sign indefinite stabilizing solutions of (1.1) and (1.4). Nevertheless, if they exist, the stabilizing solutions of these Riccati equations are always unique [19]. The following

* Received by the editors July 2, 1990; accepted for publication (in revised form) March 27, 1991.

† Code 39103, Naval Weapons Center, China Lake, California 93555.

variant of Theorem 3.3 unites all of these Riccati equations and provides an eigenvalue characterization of the nonsingular sign indefinite stabilizing solutions, as well as a test criterion that can be used to determine a lower bound for the γ semi-infinite interval of existence for $X(\gamma)$. A positive definite unique stabilizing solution of (1.1) exists on some γ half-line if and only if (i) an antistabilizing negative definite minimal solution Z_- of (1.2) exists, (ii) the negative semidefinite maximal antistabilizing solution $W_-(\gamma)$ of (1.3) exists on the same γ half-line, and (iii) the spectral radius of $-W_-(\gamma)Z_-^{-1}$ is less than unity. This spectral radius test criterion is apparently new and is one of the main results of this paper.

The results are more delicate for positive semidefinite stabilizing solutions $X(\gamma)$ of (1.1) defined on some half-line, since the antistabilizing solution of FARE need not exist [7]. Nevertheless, when $X(\gamma)$ exists, by a fundamental theorem about the regularity of solutions of (1.1) summarized in the Appendix, the “slightly” perturbed game Riccati equation for some sufficiently small positive real parameter ε

$$(1.5) \quad A^T X^\varepsilon + X^\varepsilon A + X^\varepsilon \left(\frac{1}{\gamma^2} B_1 B_1^T - B_2 B_2^T \right) X^\varepsilon + C^T C + \varepsilon I = 0$$

always has a positive definite stabilizing solution $X^\varepsilon(\gamma)$ defined on the same γ -half-line. Since the matrix norm of the difference between $X^\varepsilon(\gamma)$ and $X(\gamma)$ tends to zero as $\varepsilon \downarrow 0$, the spectral radius test in Theorem 3.3 is reformulated for the positive semidefinite stabilizing solution $X(\gamma)$ by utilizing the “nearby” solutions $X^\varepsilon(\gamma)$.

The game Riccati equations required for a full solution to the H_∞ -control problem are more general than (1.1) and (1.4), which are studied in the first two sections. However, all of the assumptions required in the main theorems and lemmas in those sections are weak enough such that they are directly applicable to the more general Riccati equations. Actually, a full solution to the H_∞ -control problem requires two Riccati equations, which are

$$(1.6) \quad \begin{aligned} & (A - B_2 E_1^{-1} D_{12}^T C_1)^T \hat{X} + \hat{X} (A - B_2 E_1^{-1} D_{12}^T C_1) \\ & + \hat{X} \left(\frac{1}{\gamma^2} B_1 B_1^T - B_2 E_1^{-1} B_2^T \right) \hat{X} \\ & + C_1^T (I - D_{12} E_1^{-1} D_{12}^T) C_1 = 0, \end{aligned}$$

$$(1.7) \quad \begin{aligned} & (A - B_1 D_{21}^T E_2^{-1} C_2) \hat{Y} + \hat{Y} (A - B_1 D_{21}^T E_2^{-1} C_2)^T \\ & + \hat{Y} \left(\frac{1}{\gamma^2} C_1^T C_1 - C_2^T E_2^{-1} C_2 \right) \hat{Y} \\ & + B_1 (I - D_{21}^T E_2^{-1} D_{21}) B_1^T = 0. \end{aligned}$$

Here the superscript -1 denotes the matrix inverse, and I is the $n \times n$ identity matrix. The matrices in these equations arise in a linear system of the form

$$\begin{aligned} \dot{x} &= Ax + B_1 w + B_2 u, \\ z &= C_1 x + D_{12} u, \\ y &= C_2 x + D_{21} w, \end{aligned}$$

where x is the $n \times 1$ state vector, w is the $n \times p$ disturbance input, u is the $n \times m$ control input, z is the $q \times n$ error output, and y is the $r \times n$ measured output. Related variants of these state equations are described in Safonov and Limebeer [22] and Zhou and Khargonekar [27]. To ensure that the H_∞ -control problem has a solution, Petersen,

Anderson, and Jonckheere [18] require that this system satisfy the following additional assumptions (see also Glover and Doyle [6] and Bernstein and Haddad [1]): $D_{12}^T D_{12} = E_1 > 0$, $D_{21} D_{21}^T = E_2 > 0$,

$$(1.8) \quad \text{rank} \begin{bmatrix} A - j\omega I & B_2 \\ C_1 & D_{12} \end{bmatrix} = n + m \quad \text{for all } \omega \geq 0,$$

$$(1.9) \quad \text{rank} \begin{bmatrix} A - j\omega I & B_1 \\ C_2 & D_{21} \end{bmatrix} = n + r \quad \text{for all } \omega \geq 0.$$

To qualify for the H_∞ -control problem, both game Riccati equations (1.6) and (1.7) must have positive semidefinite stabilizing solutions $X_\infty(\gamma)$ and $Y_\infty(\gamma)$ on a common γ interval of existence and satisfy the spectral inequality $\rho(X_\infty(\gamma)Y_\infty(\gamma)) < \gamma^2$ [6].

Throughout the paper, the γ -dependent monotonic ordering of the maximal and minimal solutions of the game Riccati equations on the γ -half-line is established. These monotonicity results are extended to give a monotonicity result for the spectral radius $\rho(X_\infty(\gamma)Y_\infty(\gamma))$. In the final section, a complete description of the rank of the stabilizing solution of the game Riccati equation is presented, including the fact that the rank is independent of γ . While neither of these apparently new results are unexpected, nevertheless, they do further characterize the γ partial ordering and γ -independent rank properties of these solutions.

2. Stabilizing and antistabilizing solutions of the dual game Riccati equation. In this section a formula is derived that expresses the maximal stabilizing solution of the dual game Riccati equation as the sum of the maximal stabilizing solution of the FARE and the minimal stabilizing solution of the BRRE. A complementary formula for the minimal antistabilizing solution of the dual game Riccati equation is also included. Using these formulas a simple eigenvalue test determines the inertia of either solution.

Before deriving these results, these additional concepts are introduced. The *inertia* of the matrix \hat{M} is the triple $\text{In}(\hat{M}) = (\nu(\hat{M}), \delta(\hat{M}), \pi(\hat{M}))$, where $\nu(\hat{M})$, $\delta(\hat{M})$, $\pi(\hat{M})$ are, respectively, the number of eigenvalues of \hat{M} counting multiplicities with negative, zero, and positive real parts [14]. The symmetric matrix \hat{M} is *nonsingular* if $\delta(\hat{M}) = 0$. An $n \times n$ symmetric matrix \hat{M} for $n \geq 2$ is *sign indefinite* if $\text{In}(\hat{M}) = (\nu(\hat{M}), 0, \pi(\hat{M}))$ with $\nu(\hat{M}) > 0$ and $\pi(\hat{M}) > 0$. Here and elsewhere, $X > P(X \geq P)$ denotes the usual partial order for symmetric matrices and means that $X - P$ is positive (semi)definite [10].

Recall that a real symmetric solution P of the algebraic Riccati equation

$$(2.1) \quad A^T P + PA - PMP + Q = 0$$

with real symmetric matrices M and Q , $M \geq 0$ is a *maximal (minimal)* [7], [21] solution if $P \geq \hat{P}$ ($P \leq \hat{P}$) for any other symmetric solution \hat{P} of (2.1). Whenever they exist, the *maximal and minimal* solutions will be denoted by P_+ and P_- , respectively. Independent of the inertia of M , a real symmetric matrix P that satisfies (1.5) is said to be *stabilizing* if $A - MP$ is stable. The matrix $A - MP$ is *stable* if the real part of each eigenvalue of $(A - MP)$ is less than zero (i.e., $\text{Re } \lambda_i(A - MP) < 0$, $i = 1, \dots, n$). If all of the eigenvalues of $A - MP$ satisfy the inequality $\text{Re } \lambda_i(A - MP) \leq 0$, then P is a *strong solution*. P is an *antistabilizing* solution if $-(A - MP)$ is stable. The spectral radius [10] of an $n \times n$ matrix M is the maximum modulus over the set of all eigenvalues $\rho(M) = \max_{1 \leq j \leq n} |\lambda_j(M)|$. The 2-norm of M is $\|M\|_2 = [\rho(M^T M)]^{1/2}$ and is the maximum singular value of M .

Associated with the algebraic Riccati equation (1.5) is the $2n \times 2n$ Hamiltonian matrix

$$(2.2) \quad H = \begin{pmatrix} A & -M \\ -Q & -A^T \end{pmatrix}.$$

If P is any solution of (2.1), then $(P, -I)H \begin{pmatrix} I \\ P \end{pmatrix} = 0$. The notation $\text{Ric}(P, H) = 0$ will symbolize this relation between a solution of the Riccati equation (2.1) and the Hamiltonian (2.2). The Hamiltonians for the game Riccati equation (1.1) and the filter algebraic Riccati equation are, respectively,

$$H_\infty(\gamma) = \begin{pmatrix} A & (1/\gamma^2)B_1B_1^T - B_2B_2^T \\ -C^TC & -A^T \end{pmatrix}, \quad H_2 = \begin{pmatrix} A^T & -C^TC \\ -B_2B_2^T & -A \end{pmatrix}.$$

Whenever the solutions Z_+ or Z_- for FARE exist, define the Hamiltonian for BRRE as

$$H(Z_\pm, \gamma) = \begin{pmatrix} (A - Z_\pm C^TC)^T & -C^TC \\ (1/\gamma^2)B_1B_1^T & -(A - Z_\pm C^TC) \end{pmatrix}.$$

An eigenvalue λ of A is (A, B) controllable if $\text{rank}[A - \lambda I, B] = n$. The pair (A, B) is stabilizable if and only if (A, B) has no uncontrollable eigenvalue in the closed right half plane. Observability and detectability follow by duality.

Recently, Petersen, Anderson, and Jonckheere [18] have proved the strict bounded real lemma for nonminimal realizations. For completeness, their lemma will be repeated here with an added statement and proof about the minimality of the stabilizing solution. Willems [25] has proved, again for A stable, that if the bounded real Riccati equation admits a real symmetric solution, then $\|G(s)\|_\infty < 1$.

LEMMA 2.1 (strict bounded real lemma). *The following are equivalent:*

- (i) A is stable and the transfer matrix $G(s) = C(sI - A)^{-1}B$ evaluated on the imaginary axis satisfies the inequality $\|G(s)\|_\infty = \max_{\omega \in \mathbb{R}} \|G(j\omega)\|_2 < 1$,
- (ii) The Riccati equation

$$(2.3) \quad A^TP + PA + PBB^TP + C^TC = 0$$

has a stabilizing solution $P_S \geq 0$. Moreover, the solution P_S is unique and minimal.

Proof. Suppose that P is any other real symmetric solution of (2.2). After solving (2.3) for C^TC , substitute its value in the Riccati equation for P_S . After arranging the two solutions, it follows that their difference satisfies the Lyapunov equation

$$(2.4) \quad \begin{aligned} & (A + BB^TP_S)(P_S - P) + (P_S - P)(A + BB^TP_S) \\ & - (P_S - P)BB^T(P_S - P) = 0 \end{aligned}$$

Since $A + BB^TP_S$ is stable, the conclusion follows by Lyapunov inertia theory [14, p. 447]. \square

Bounded real Riccati equations satisfy monotonicity and extension properties that are derivable from general theorems of this type in Ran and Vreugdenhil [21], or Gohberg, Lancaster, and Rodman [7] and for monotonicity Wimmer [26]. The following result is found in Petersen, Anderson, and Jonckheere [18].

THEOREM 2.2 (bounded real extension theorem). *Suppose that A is stable, \hat{Q} is a real symmetric matrix, and the Riccati equation $A^TP + PA + PBB^TP + \hat{Q} = 0$ has a real symmetric solution \hat{P} . Furthermore, suppose that $\hat{Q} \geq Q \geq 0$. Then the Riccati equation $A^TP + PA + PBB^TP + Q = 0$ will have a unique strong solution $P \geq 0$ such that $P \leq \hat{P}$.*

Another key result is the relationship given by Boyd, Balakrishnan, and Kabamba (BBK) [2] that connects the singular values of a stable transfer matrix $G(s) = C(sI - A)^{-1}B$ and the spectrum of the Hamiltonian matrix

$$H(\gamma) = \begin{pmatrix} A & (1/\gamma^2)BB^T \\ -C^TC & -A^T \end{pmatrix}.$$

Their theorem does not require any special observability or controllability conditions on the system $\{A, B, C\}$.

THEOREM 2.3 (BBK theorem). *Suppose that A is stable. $\|G(s)\|_\infty < \lambda$ if and only if $\delta(H(\gamma)) = 0$.*

The maximal stabilizing solution of FARE and the minimal or maximal stabilizing solution of BRRE—depending on whether (1.3) or (2.3) is solved—are now combined to define the maximal and minimal stabilizing solution of the dual game Riccati equation.

THEOREM 2.4. *If the stabilizing Z_+ of FARE exists and if $\delta(H(Z_+, \hat{\gamma})) = 0$ for some $\hat{\gamma}$, $0 < \hat{\gamma} < \infty$, then the unique maximal stabilizing solution $W_+(\gamma)$ of $\text{Ric}(W, H(Z_+, \gamma)) = 0$ exists for all γ , $0 < \hat{\gamma} \leq \gamma \leq \infty$. Furthermore, the stabilizing solution $Y_+(\gamma)$ of $\text{Ric}(Y, H_\infty^T(\gamma)) = 0$ also exists on the same interval and $Y_+(\gamma) = Z_+ + W_+(\gamma)$, $0 < \hat{\gamma} \leq \gamma \leq \infty$. Moreover, if $0 < \hat{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \infty$, then $Y_+(\gamma_1) \leq Y_+(\gamma_2)$.*

Proof. Since $\delta(H(Z_+, \hat{\gamma})) = 0$ and $A - Z_+C^TC$ is stable, the transfer matrix $G_1(s)$ for the realization $(A - Z_+C^TC, B_1, C)$ satisfies the inequality $\|G_1(s)\|_\infty < \hat{\gamma}$ (Theorem 2.3). Let $W = -P$ in $\text{Ric}(W, H(Z_+, \gamma)) = 0$ and obtain the equation

$$(2.5) \quad (A - Z_+C^TC)P + P(A - Z_+C^TC)^T + PC^TCP + \frac{1}{\gamma^2} B_1 B_1^T = 0$$

By Lemma 2.1, the unique minimal solution $P_-(\hat{\gamma}) \geq 0$ of (2.5) exists.

The explicit Riccati equation for $\text{Ric}(W, H(Z_+, \gamma)) = 0$ is

$$(2.6) \quad (A - Z_+C^TC)W + W(A - Z_+C^TC)^T - WC^TCW - \frac{1}{\gamma^2} B_1 B_1^T = 0.$$

Transforming back the maximal stabilizing solution $W_+(\hat{\gamma}) = -P_-(\hat{\gamma}) \leq 0$ of (2.6) is obtained.

Next, we show that $W_+(\gamma)$ exists for any γ in the infinite interval $0 < \hat{\gamma} \leq \gamma \leq \infty$ and satisfies the inequality

$$(2.7) \quad W_+(\hat{\gamma}) \leq W_+(\gamma), \quad 0 < \hat{\gamma} \leq \gamma \leq \infty.$$

Since $A - Z_+C^TC$ is stable for any γ and $(1/\gamma^2)B_1 B_1^T \leq (1/\hat{\gamma})B_1 B_1^T$, the strong solution $P_-(\gamma)$ of (2.5) exists and satisfies the partial order $0 \leq P_-(\hat{\gamma}) \leq P_-(\gamma)$ by Theorem 2.2. To show that $W_+(\gamma)$ is stabilizing, note that $(1/\gamma)\|G_1(s)\|_\infty \leq (1/\hat{\gamma})\|G_1(s)\|_\infty < 1$ and then apply Theorem 2.3 to show that $\delta(H(Z_+, \gamma)) = 0$ for all $\gamma \geq \hat{\gamma}$.

The next step is to show that the sum $Y_+(\gamma) = Z_+ + W_+(\gamma)$ is a solution of (1.4). First, substitute it into (1.4) and rearrange the equation to obtain the identity (the subscripts are unnecessary)

$$(2.8) \quad \begin{aligned} & A(Z + W(\gamma)) + (Z + W(\gamma))A^T - (Z + W(\gamma))C^TC(Z + W(\gamma)) \\ & \quad + B_2 B_2^T - \frac{1}{\gamma^2} B_1 B_1^T \\ & = (AZ + ZA^T - ZC^TCZ + B_2 B_2^T) + (A - ZC^TC)W(\gamma) \\ & \quad + W(\gamma)(A - ZC^TC)^T - W(\gamma)C^TCW(\gamma) - \frac{1}{\gamma^2} B_1 B_1^T = 0. \end{aligned}$$

Thus, the sum of any two such solutions is a solution of the dual game Riccati equation. Since $(A - Z_+ C^T C) + P_-(\gamma) C^T C$ is stable, the sum $Z_+ + W_+(\gamma)$ is stabilizing. The maximality of $Y_+(\gamma)$ can be established by the Lyapunov arguments used in the proof of Lemma 2.1. By (2.7), $Y_+(\gamma_1) - Y_+(\hat{\gamma}_2) = W_+(\gamma_1) - W_+(\hat{\gamma}_2) \geq 0$, which establishes the partial order of $Y_+(\gamma)$. \square

The properties of the dual minimal antistabilizing solution $Y_-(\gamma)$ will now be established by similar arguments applied to the Riccati equation $\text{Ric}(W, -H(Z_-, \gamma)) = 0$.

THEOREM 2.5. *If the antistabilizing Z_- of FARE exists and if $\delta(H(Z_-, \hat{\gamma})) = 0$ for some $\hat{\gamma}$, $0 < \hat{\gamma} < \infty$, then the unique minimal stabilizing solution $W_-(\gamma)$ of $\text{Ric}(W, -H(Z_-, \gamma)) = 0$ exists for all γ , $0 < \hat{\gamma} \leq \gamma \leq \infty$. Furthermore, the antistabilizing solution $Y_-(\gamma)$ of $\text{Ric}(Y, H_\infty^T(\gamma)) = 0$ also exists on the same interval and the relation is valid $Y_-(\gamma) = Z_- + W_-(\gamma)$, $0 < \hat{\gamma} \leq \gamma \leq \infty$. Moreover, if $0 < \hat{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \infty$, then $Y_-(\gamma_2) \leq Y_-(\gamma_1)$.*

Proof. The congruence transformation

$$\begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix} H(Z_-, \hat{\gamma}) \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} = -H^T(Z_-, \hat{\gamma})$$

implies by Sylvester's law of inertia [10] that $\delta(-H(Z_-, \hat{\gamma})) = \delta(-H^T(Z_-, \hat{\gamma})) = \delta(H(Z_-, \hat{\gamma})) = 0$.

Since $A - Z_- C^T C$ is antistable, the transfer matrix $\hat{G}(s)$ for the realization $-(A - Z_- C^T C), B_1, C$ satisfies the inequality $\|\hat{G}(s)\|_\infty < \hat{\gamma}$ by Theorem 2.3. Invoking the reasoning in Theorem 2.4, $\delta(-H^T(Z_-, \gamma)) = 0$ for all $\gamma \geq \hat{\gamma}$.

The explicit Riccati equation for $\text{Ric}(W, -H(Z_-, \gamma)) = 0$ is

$$(2.9) \quad [-(A - Z_- C^T C)]W + W[-(A - Z_- C^T C)]^T + WC^T C W + \frac{1}{\gamma^2} B_1 B_1^T = 0.$$

Again by Lemma 2.1 the stabilizing solution $W_-(\gamma)$ of (2.9) exists. The extension and solution monotonicity properties that are derived in Theorem 2.1 can be used to show that $W_-(\gamma)$ exists on the semi-infinite interval and satisfies the partial order

$$(2.10) \quad W_-(\gamma) \leq W_-(\hat{\gamma}), \quad 0 < \hat{\gamma} \leq \gamma \leq \infty.$$

The sum $Y_-(\gamma) = Z_- + W_-(\gamma)$ is a solution of $\text{Ric}(Y, H_\infty^T(\gamma)) = 0$ because it satisfies (2.8). Since $-(A - Z_- C^T C) + W_-(\gamma) C^T C$ is stable, it follows that $W_-(\gamma)$ is antistabilizing and that the sum $Y_-(\gamma)$ is antistabilizing. By (2.10), $Y_-(\gamma) - Y_-(\hat{\gamma}) = W_-(\gamma) - W_-(\hat{\gamma}) \geq 0$, thus, the partial order of $Y_+(\gamma)$ is established. \square

Using the decomposition theorems just derived, the inertia properties of the antistabilizing and stabilizing solutions of (1.4) can be easily derived. Moreover, they clearly reveal how and why sign indefinite stabilizing solutions can occur. The existence of sign indefinite stabilizing solutions of game Riccati equations for $0 < \gamma < \infty$ highlights a fundamental difference between them and the stabilizing solutions of the standard control, filtering or the bounded real Riccati equations, which cannot be sign indefinite. Only the inertia properties of the stabilizing solutions will be proved.

Let $\text{Ker}(M)$ denote the null space or kernel of the matrix M and $\text{Im}(M)$ denote the range or image of M .

THEOREM 2.6. *If the nonzero solution Z_+ of $\text{Ric}(Z, H_2) = 0$ is nonsingular and $W_+(\gamma)$ of $\text{Ric}(W, H(Z_+, \gamma)) = 0$ exists for some $\hat{\gamma}$, $0 < \hat{\gamma} < \infty$, then the solution $Y_+(\gamma)$ of $\text{Ric}(Y, H_\infty^T(\gamma)) = 0$ exists for all γ $0 < \hat{\gamma} \leq \gamma \leq \infty$ —by Theorem 2.4—and satisfies the following properties:*

$$(a) \quad \text{In}(Y_+(\gamma)) = \text{In}(I + W_+(\gamma)Z_+^{-1}),$$

- (b) $\delta(Y_+(\gamma)) = 0$ if and only if $\lambda_i(-W_+(\gamma)Z_+^{-1}) \neq 1$, $i = 1, \dots, n$,
(c) $Y_+(\gamma) > 0$ for all γ , $0 < \hat{\gamma} \leq \gamma \leq \infty$ if and only if $\rho(-W_+(\hat{\gamma})Z_+^{-1}) < 1$.

Proof. Since Z_+ is nonsingular by Sylvester's law of inertia $\text{In}(Y_+(\gamma)) = \text{In}(Y_+(\gamma)Z_+^{-1}) = \text{In}(I + W_+(\gamma)Z_+^{-1})$ and (a)–(b) follow from this identity. Since $Z_+ > 0$ and $W_+(\gamma) \leq 0$, the spectral radius test [10], is equivalent to the statement $Y_+(\hat{\gamma}) > 0$ if and only if $\rho(-W_+(\hat{\gamma})Z_+^{-1}) < 1$. The inverse square root $Z_+^{-1/2}$ exists, because $Z_+ > 0$. Now, by (2.7), $W_+(\hat{\gamma}) \geq W_+(\gamma)$ for $0 < \hat{\gamma} \leq \gamma \leq \infty$. The spectral sets for $-W_+(\gamma)Z_+^{-1}$ and $-Z_+^{-1/2}W_+(\gamma)Z_+^{-1/2}$ are identical and they only contain real nonnegative eigenvalues [10, p. 468]. Thus, for $0 < \hat{\gamma} \leq \gamma \leq \infty$,

$$\begin{aligned} \rho(-W_+(\gamma)Z_+^{-1}) &= \rho(-Z_+^{-1/2}W_+(\gamma)Z_+^{-1/2}) \\ &= \|(Z_+^{-1/2}W_+(\gamma)Z_+^{-1/2})\| \leq \|(Z_+^{-1/2}W_+(\hat{\gamma})Z_+^{-1/2})\| \\ &= \rho(-Z_+^{-1/2}W_+(\hat{\gamma})Z_+^{-1/2}) = \rho(-W_+(\hat{\gamma})Z_+^{-1}). \quad \square \end{aligned}$$

In fact, if $Y_+(\hat{\gamma})$ is sign indefinite for some $\hat{\gamma}$ and if $Y_+(\gamma)$ exists on the half-line $0 < \hat{\gamma} \leq \gamma \leq \infty$, then by the continuity of eigenvalues and by the monotone nondecreasing property of $Y_+(\gamma)$ as a function of γ , it will be singular for some finite value of γ because the $\text{In}(Y_+(\gamma))$ depends continuously on the eigenvalues of $Y_+(\gamma)$ and $Y_+(\infty) = Z_+$.

The following converse of Theorem 2.4 is now obtained.

THEOREM 2.7. *If there exists a maximal stabilizing solution $Y_+(\gamma)$ of $\text{Ric}(Y, H_\infty^T(\gamma)) = 0$ on some γ semi-infinite interval $0 < \hat{\gamma} \leq \gamma \leq \infty$, then the following solutions exist on the same interval:*

- (i) *the stabilizing solution Z_+ of $\text{Ric}(Z, H_2) = 0$ exists and $\delta(H(Z_+, \gamma)) = 0$,*
(ii) *the stabilizing solution $W_+(\gamma)$ of $\text{Ric}(W, H(Z_+, \gamma)) = 0$ exists.*

Proof. Since $Y_+(\gamma)$ exists at $\gamma = \infty$, the solution $Z_+ = Y_+(\infty)$ clearly exists. Since Z_+ exists, the similarity transformation

$$(2.11) \quad \begin{pmatrix} I & 0 \\ -Z_+ & I \end{pmatrix} H_\infty^T(\gamma) \begin{pmatrix} I & 0 \\ Z_+ & I \end{pmatrix} = H(Z_+, \gamma).$$

shows that $\text{In}(H_\infty^T(\gamma)) = \text{In}(H(Z_+, \gamma))$. Thus, $\delta(H(Z_+, \gamma)) = 0$ because $Y_+(\gamma)$ is stabilizing now, the proof in Theorem 2.4 shows that $W_+(\gamma)$ exists on the same interval. \square

By combining the proof of Theorems 2.5 and 2.6, the following dual result for the antistabilizing solutions is obtained.

THEOREM 2.8. *If there exists a minimal antistabilizing solution $Y_-(\gamma)$ of $\text{Ric}(Y, H_\infty^T(\gamma)) = 0$ on some γ semi-infinite interval $0 < \hat{\gamma} \leq \gamma \leq \infty$, then on the same interval*

- (i) *the antistabilizing solution Z_- of $\text{Ric}(Z, H_2) = 0$,*
(ii) *the stabilizing solution $W_-(\gamma)$ of $\text{Ric}(W, -H(Z_-, \gamma)) = 0$ exists.*

Proof. Since $Y_-(\gamma)$ exists at $\gamma = \infty$, the solution $Z_- = Y_-(\infty)$ clearly exists. Since (2.11) is a similarity transformation $\text{In}(H_\infty^T(\gamma)) = \text{In}(H(Z_\pm, \gamma))$, which means that $\delta(H(Z_-, \gamma)) = 0$ because $Y_-(\gamma)$ is antistabilizing. Now the proof in Theorem 2.5 can be applied to show that $W_-(\gamma)$ exists. \square

3. Stabilizing solutions of the game Riccati equation. In this section, the stabilizing solution $X(\gamma)$ of the game Riccati control equation (1.1) is related via the separation formulas derived in the previous section to the antistabilizing solutions Z_- and $W_-(\gamma)$ that are factors of the dual game Riccati equation antistabilizing solution. When $\gamma = \infty$

and $X(\gamma)$ is nonsingular, then this relationship reduces to the familiar fact, namely, $Z_- = -X^{-1}(\infty)$ [11].

To keep the controllability and observability assumptions to a minimum in the subsequent theorems, some properties of the stabilizing solutions $X(\gamma)$ of (1.1) are deduced. However, before these deductions are discussed, a key lemma that avoids assuming that the pair (C, A) is observable is introduced. This is especially important for the Riccati equations described by (1.6) and (1.7). Then a preliminary theorem is proved that investigates the role of stabilizability and detectability, when $X(\gamma)$ is a stabilizing solution.

This decomposition is used in Petersen, Anderson, and Jonckheere [18] and Hinrichsen and Pritchard [9]. It is a consequence of two key facts. Namely, the linear subspaces determined by $\text{Im } X(\gamma)$ and $\text{Ker } X(\gamma)$ are orthogonal and $\text{Ker } X(\gamma)$ is an A -invariant subspace of $\text{Ker } C$.

LEMMA 3.1. *Let X be any real symmetric solution of $\text{Ric}(X, H_\infty(\gamma)) = 0$. The matrices A , B_1 , B_2 , and C with respect to any orthonormal basis compatible with the decomposition $R^n = \text{Im } X \oplus \text{Ker } X$ can be written as*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad C = [C_{11}, 0], \quad B_1 = \begin{bmatrix} B_{11} \\ B_{22} \end{bmatrix}, \quad B_2 = \begin{bmatrix} B_{21} \\ B_{22} \end{bmatrix}, \quad X = \begin{bmatrix} X_{11} & 0 \\ 0 & 0 \end{bmatrix}.$$

Moreover, if X is not identically zero, $\delta(X_{11}) = 0$ [9], and X_{11} satisfies the equation

$$(3.1) \quad A_{11}^T X_{11} + X_{11} A_{11} + X_{11} \left(\frac{1}{\gamma^2} B_{11} B_{11}^T - B_{21} B_{21}^T \right) X_{11} + C_{11}^T C_{11} = 0.$$

Some of the deductions in the next theorem are well known for $\text{Ric}(X, H_\infty(\infty)) = 0$, and (iii) would follow from the H infinity papers such as [4] or [18]. However, $\text{Ric}(X, H_\infty(\gamma)) = 0$ can have sign indefinite stabilizing solutions. Because local properties of stabilizing solutions are important for numerical studies [12], a proof is included. In fact, the next theorem, when combined with the result in the Appendix and with the metric topology introduced by Gahinet and Laub [5], can be used to justify the informal statement: the neighbourhoods defined by “small admissible perturbations” of $\text{Ric}(X, H_\infty(\gamma)) = 0$ will always contain unique stabilizing solutions, whenever the unperturbed equation has a stabilizing solution. Subsequently, a theorem will show that if a positive semidefinite stabilizing solution exists for a single γ , $0 < \gamma < \infty$ then it exists for any larger value of γ . This extension property and the minimal assumptions that guarantee it are apparently new.

Let $H_\infty^\varepsilon(\gamma)$ denote the $2n \times 2n$ Hamiltonian matrix for the slightly perturbed game equation (1.8) obtained from $H_\infty(\gamma)$ by perturbing the $(2, 1)$ block matrix $C^T C$ by εI for $\varepsilon > 0$ and $H^\varepsilon(Z_\pm, \gamma)$ is the $2n \times 2n$ matrix obtained from $H(Z_\pm, \gamma)$ by perturbing the $(1, 2)$ block matrix $C^T C$ by εI for $\varepsilon > 0$.

THEOREM 3.2. *Suppose (that $\text{Ric}(X, H_\infty(\gamma)) = 0$ has a nonzero stabilizing solution $X(\gamma)$ for some γ , $0 < \gamma < \infty$, then*

- (i) $\delta(H_2) = 0$, otherwise,
- (ii) if $X(\gamma) \geq 0$, then the pair (A, B_2) is stabilizable,
- (iii) if $X(\gamma) > 0$, then the pair $(C, -A)$ is detectable.

Proof. Since $X(\gamma)$ is stabilizing, the closed-loop matrix $A + ((1/\gamma^2)B_1 B_1^T - B_2 B_2^T)X(\gamma)$ is stable. Since the decomposition in Lemma 3.1 can be realized by a real orthogonal similarity transformation, the A_{22} block of the transformed closed loop matrix will be stable. Now the dimension of X_{11} equals the dimension of the block matrix A_{11} . The unstable eigenvalues (if any) of A_{11} are clearly $(A_{11}, (1/\gamma^2)B_1 B_1^T - B_2 B_2^T)$ -controllable. Since $X_{11}(\gamma)$ is nonsingular, (3.1) can be

rearranged to yield the equivalent equation

$$(3.2) \quad \begin{aligned} A_{11} + \left(\frac{1}{\gamma^2} B_{11} B_{11}^T - B_{21} B_{21}^T \right) X_{11}(\gamma) \\ = -X_{11}^{-1}(\gamma) (A_{11} + X_{11}^{-1}(\gamma) C_{11}^T C_{11})^T X_{11}(\gamma) \end{aligned}$$

Equation (3.2) clearly shows that $(A_{11} + X_{11}^{-1}(\gamma) C_{11}^T C_{11})$ is antistable, and so the stable or pure imaginary eigenvalue of A_{11} (if any) are clearly $(A_{11}, C_{11}^T C_{11})$ -observable. Equation (3.2) shows that the purely imaginary eigenvalues (if any) of A are (C, A) -observable.

By Theorem A in the Appendix, the perturbed Riccati equation $\text{Ric}(X^\varepsilon, H_\infty^\varepsilon(\gamma)) = 0$ will have a nonsingular stabilizing solution $X^\varepsilon(\gamma)$. Rearranging the equation, it becomes

$$(3.3) \quad \begin{aligned} (A - B_2 B_2^T X^\varepsilon(\gamma))^T X^\varepsilon(\gamma) + X^\varepsilon(\gamma) (A - B_2 B_2^T X^\varepsilon(\gamma)) \\ + X^\varepsilon(\gamma) \left(\frac{1}{\gamma^2} B_1 B_1^T + B_2 B_2^T \right) X^\varepsilon(\gamma) + C^T C + \varepsilon I = 0. \end{aligned}$$

Since $X^\varepsilon(\gamma)$ is nonsingular and $(\varepsilon I, A - B_2 B_2^T X^\varepsilon(\gamma))$ is observable, it follows by Lyapunov inertia theory [14, p. 448], that $\text{In}(A - B_2 B_2^T X^\varepsilon(\gamma)) = \text{In}(-X^\varepsilon(\gamma))$ and $\delta(A - B_2 B_2^T X^\varepsilon(\gamma)) = 0$. At the very least, all of the eigenvalues of A on the imaginary axis are (A, B_2) -controllable and so $\delta(H_2) = 0$ [13], [15], otherwise if $X(\gamma) \geq 0$, then $X^\varepsilon(\gamma) > 0$ (Appendix) and so (A, B_2) is stabilizable. The second claim follows by a similarity transformation identical to (3.2). \square

The following theorem establishes the existence of nonsingular stabilizing solutions of the control Riccati equation (1.1) by a Hamiltonian and spectral radius test over the parameter interval. These tests can be implemented by a Hamiltonian bisection search [9], [2] or by a gradient search on the spectral radius [17]. Moreover, these conditions completely characterize the nonsingular stabilizing solutions of (1.1), including the sign indefinite ones.

THEOREM 3.3. *The nonsingular stabilizing solution $X(\gamma)$ of $\text{Ric}(X, H_\infty(\gamma)) = 0$ exists on the semi-infinite interval $0 < \hat{\gamma} \leq \gamma \leq \infty$ if and only if*

- (i) *the equation $\text{Ric}(Z, H_2) = 0$ has a minimal antistabilizing solution $Z_- < 0$,*
- (ii) *the stabilizing solution $W_-(\gamma)$ of $\text{Ric}(W, -(H(Z_-, \gamma))) = 0$ exists on the same interval,*
- (iii) *the real eigenvalues of the matrix $W_-(\gamma) Z_1^{-1}$ satisfy the inequality $\lambda_i(-W_-(\gamma) Z_1^{-1}) \neq 1$ on the same interval.*

Furthermore, $\text{In}(X(\gamma)) = \text{In}(I + W_-(\gamma) Z_1^{-1})$ and $X(\gamma) > 0$ if and only if $\rho(W_-(\gamma)(-Z_1^{-1})) < 1$.

Proof. First, the sufficiency conditions are assumed. The matrix $Y_-(\gamma) = Z_- + W_-(\gamma)$ is an antistabilizing solution of the dual game Riccati equation by Theorem 2.5. Clearly, if the real eigenvalues of $W_-(\gamma) Z_1^{-1}$ satisfy (iii) then $Y_-(\gamma)$ is nonsingular by Theorem 2.6. If $Y_-(\gamma)$ is nonsingular, then its inverse will satisfy the game Riccati equation (1.1) and, by an obvious modification of (3.2), it will be the unique stabilizing solution of (1.1), too. Since the stabilizing solution of (1.1) is unique $X(\gamma) = -(Y_-(\gamma))^{-1}$ so $X(\gamma)$ has the asserted properties. Z_- is nonsingular by hypothesis and so by Theorem 2.6 $\text{In}(Y_-(\gamma)) = \text{In}(I + W_-(\gamma) Z_1^{-1})$.

If $X(\gamma)$ is the nonsingular stabilizing solution on $0 < \hat{\gamma} \leq \gamma \leq \infty$, then $-X^{-1}(\infty)$ is an antistabilizing solution of $\text{Ric}(Z, H_2) = 0$ and so (i) is satisfied. Again, by (3.2), for any γ in the interval $Y_-(\gamma) = -X^{-1}(\gamma)$ is an antistabilizing solution of $\text{Ric}(W, H_\infty^T(\gamma)) = 0$ on the same interval. By Theorems 2.5 and 2.8, (ii) and (iii) are satisfied. \square

When the stabilizing solution is semidefinite, then (i), (ii), and (iii) are satisfied by the perturbed Riccati equation $\text{Ric}(W, H_\infty^\varepsilon(\gamma)^T) = 0$ and, conversely, these conditions guarantee the existence of a strong solution. A strong solution of (1.1) will also be stabilizing if either the assumption $\delta(H_\infty(\gamma) = 0$ is satisfied on some semi-infinite interval or the less familiar assumption (1.8) is satisfied.

THEOREM 3.4. *If the stabilizing solution $X(\gamma) \geq 0$ of $\text{Ric}(X, H_\infty(\gamma)) = 0$ exists on the semi-infinite interval $0 < \hat{\gamma} \leq \gamma \leq \infty$, then for all sufficiently small $\varepsilon > 0$*

(i) *the equation $\text{Ric}(Z, (H_\infty^\varepsilon(\infty))^T) = 0$ has a minimal solution $Z_-^\varepsilon < 0$ on the same interval,*

(ii) *the stabilizing solution $W_-^\varepsilon(\gamma)$ of $\text{Ric}(W, -(H^\varepsilon(Z_-, \gamma))) = 0$ exists on the same interval,*

(iii) *the real eigenvalues of the matrix $W_-^\varepsilon(\gamma)Z_-^{\varepsilon-1}$ satisfy the inequality $\rho(-W_-^\varepsilon(\gamma)Z_-^{\varepsilon-1}) < 1$ on the same interval.*

Proof. Since $X(\gamma)$ is stabilizing and $X(\gamma) \geq 0$, Theorem A in the Appendix can be applied to $\text{Ric}(X^\varepsilon, H_\infty^\varepsilon(\gamma)) = 0$, which has a stabilizing solution $X^\varepsilon(\gamma) > 0$. The existence of Z_-^ε in (i) follows, because the pair $(C^T C + \varepsilon I, A)$ is observable, and by Theorem 3.2 the pair $(-A, B_2)$ is detectable. The other conditions are now immediate from Theorem 2.8.

Conditions (i), (ii), and (iii) and Theorem 3.3 guarantee that $\text{Ric}(X^\varepsilon, H_\infty^\varepsilon(\gamma)) = 0$ has a nonsingular stabilizing solution $X^\varepsilon(\gamma) > 0$. Theorem A in the Appendix then guarantees the existence of a strong solution, and the difference $X^\varepsilon(\gamma) - X(\gamma)$ can be made arbitrarily small as ε decreases. \square

The next theorem shows that if a positive semidefinite stabilizing solution of (1.1) exists, for a single value of γ , then it will exist for all larger values of γ ; in other words, its interval of existence is connected. This extension and connectivity property of the γ parameter can be extended to any perturbed equation that satisfies Theorem A. Thus both the game Riccati equation and “nearby” perturbed solutions will also exist over the entire gamma interval, which ensures that the regularity properties of stabilizing solutions are well defined.

THEOREM 3.5. *If a nonzero stabilizing solution $X(\hat{\gamma}) \geq 0$ of $\text{Ric}(X, H_\infty(\hat{\gamma})) = 0$ exists for some $\hat{\gamma}$, $0 < \hat{\gamma} < \infty$, then $X(\gamma) \geq 0$ exists and is stabilizing on the semi-infinite interval $0 < \hat{\gamma} \leq \gamma \leq \infty$ and obeys the partial order $X(\gamma_2) \leq X(\gamma_1)$ for $0 < \hat{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \infty$.*

Proof. Since $X(\hat{\gamma}) \geq 0$ is stabilizing, there exists by Lemma 3.1 a real orthogonal matrix S such that $X(\hat{\gamma})$ and A are transformed by the congruence transformation $S^T A S$ and $S^T X(\hat{\gamma}) S$ into the block form

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad \begin{pmatrix} X_{11}(\hat{\gamma}) & 0 \\ 0 & 0 \end{pmatrix}.$$

The closed-loop matrix $A + ((1/\gamma^2)B_1 B_1^T - B_2 B_2^T)X(\gamma)$ is stable and thus the A_{22} block of the transformed closed-loop matrix will be stable. The unstable eigenvalues (if any) of A_{11} are clearly controllable. By Lemma 3.1, the dimension of $X_{11}(\hat{\gamma})$ equals that of A_{11} , and $X_{11}(\hat{\gamma}) > 0$ is a stabilizing solution of the reduced Riccati equation (3.1). By Theorems 2.6 and 2.7, the stabilizing solution $X_{11}(\gamma) > 0$ exists on the semi-infinite intervals $0 < \hat{\gamma} \leq \gamma \leq \infty$. The inertia relation $\text{In}(S^T X(\gamma) S) = \text{In}(0, \nu(0), \pi(X_{11}(\gamma)))$ and the uniqueness of stabilizing solutions of (1.1) guarantee that $X(\gamma) \geq 0$ exists on the semi-infinite interval.

By Theorem 3.4, the solution $Y_-^\varepsilon(\gamma) < 0$ exists on the same interval, and by Theorem 2.5, $Y_-^\varepsilon(\gamma_2) \leq Y_-^\varepsilon(\gamma_1)$ for $0 < \hat{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \infty$. Moreover, the proof of Theorem 3.3 shows that the solution $X^\varepsilon(\gamma)$ of $\text{Ric}(X^\varepsilon, H_\infty^\varepsilon(\gamma)) = 0$ exists and satisfies the equation $X^\varepsilon(\gamma) = -Y_-^\varepsilon(\gamma)^{-1}$, which implies that the partial order $X^\varepsilon(\gamma_2) \leq X^\varepsilon(\gamma_1)$ is valid. \square

4. Monotonicity of the spectral radius. Since the spectral radius is not a norm, its functional properties often require special considerations. Functional properties are especially important when the spectral radius is coupled with a computational algorithm. For example Pandey *et al.* [17] search over the γ interval for the optimal H_∞ -norm by a gradient method applied to the special radius $\rho(X_\infty(\gamma)Y_\infty(\gamma))$. The next theorem shows that the discontinuities (if any) of the spectral radius $\rho(X_\infty(\gamma)Y_\infty(\gamma))$ are jumps.

THEOREM 4.1. *If there exist maximal stabilizing solutions $X_\infty(\gamma) \geq 0$ and $Y_\infty(\gamma) \geq 0$ of the respective game Riccati equations (1.6) and (1.7) defined on the same semi-infinite interval $0 < \hat{\gamma} \leq \gamma \leq \infty$, then $\rho(X_\infty(\gamma_2)Y_\infty(\gamma_2)) \leq \rho(X_\infty(\gamma_1)Y_\infty(\gamma_1))$ for $0 < \hat{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \infty$.*

Proof. Throughout the proof, γ is confined to the interval $0 < \hat{\gamma} \leq \gamma \leq \infty$. Both solutions satisfy the same γ dependent partial order $X_\infty(\gamma_2) \leq X_\infty(\gamma_1)$, $Y_\infty(\gamma_2) \leq Y_\infty(\gamma_1)$, by an obvious modification of the proof of Theorem 2.1 for the first inequality and then by invoking duality for the second inequality.

Since both solutions $X_\infty(\gamma)$ and $Y_\infty(\gamma)$ are stabilizing, Theorem A in the Appendix can be applied to the “slightly” perturbed versions of the game Riccati equations (1.6) and (1.7), which are obtained from the original equations by adding the matrix εI with $\varepsilon > 0$ to the respective matrices $C_1^T(I - D_{12}E_1^{-1}D_{12}^T)C_1$ and $B_1(I - D_{21}E_2^{-1}D_{21}^T)B_1^T$. The perturbed solutions $X_\infty^\varepsilon(\gamma) > 0$ and $Y_\infty^\varepsilon(\gamma) > 0$ inherit the same γ dependent partial order as the original solutions.

For each $\varepsilon > 0$ and γ standard balancing results [11] can be invoked to show that there exists a similarity transformation T (the explicit dependency on ε and γ is omitted) such that $X_\infty^\varepsilon(\gamma)$ and Y_∞^ε can be simultaneously diagonalized

$$T^T X_\infty^\varepsilon(\gamma) T = \Sigma_\infty^\varepsilon(\gamma) = T^{-1} Y_\infty^\varepsilon(\gamma) T^{-T}.$$

The real diagonal elements $\sigma_\infty^\varepsilon(\gamma)_i$ of the diagonal matrix $\Sigma_\infty^\varepsilon(\gamma)$ are determined by

$$\sigma_\infty^\varepsilon(\gamma)_i = (\lambda_i(X_\infty^\varepsilon(\gamma)Y_\infty^\varepsilon(\gamma)))^{1/2}, \quad i = 1, \dots, n.$$

Since $\Sigma_\infty^\varepsilon(\gamma)$ is the common unique maximal stabilizing solution of the “slightly perturbed” game Riccati equations in the transformed state equations representation, it inherits the γ dependent partial order. Thus, the spectral radius ordering

$$\rho(X_\infty^\varepsilon(\gamma_2)Y_\infty^\varepsilon(\gamma_2)) \leq \rho(X_\infty^\varepsilon(\gamma_1)Y_\infty^\varepsilon(\gamma_1))$$

for $0 < \hat{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \infty$ is established for any ε .

The eigenvalues of $X_\infty^\varepsilon(\gamma)Y_\infty^\varepsilon(\gamma)$ are all positive or zero real numbers [10, p. 468]. Furthermore, the eigenvalues of the matrices $X_\infty^{1/2}(\gamma)(X_\infty^{1/2}(\gamma)Y_\infty(\gamma))$ and $X_\infty^{1/2}(\gamma)Y_\infty(\gamma)X_\infty^{1/2}(\gamma)$ are identical. Thus, the following chain of inequalities can be obtained:

$$\begin{aligned} \rho(X_\infty(\gamma_2)Y_\infty(\gamma_2)) &= \rho(X_\infty^{1/2}(\gamma_2)Y_\infty(\gamma_2)X_\infty^{1/2}(\gamma_2)) \\ &= \|X_\infty^{1/2}(\gamma_2)Y_\infty(\gamma_2)X_\infty^{1/2}(\gamma_2)\| \\ &\leq \|X_\infty^{1/2}(\gamma_2)Y_\infty(\gamma_2)X_\infty^{1/2}(\gamma_2) - X_\infty^\varepsilon(\gamma_2)^{1/2}Y_\infty^\varepsilon(\gamma_2)X_\infty^\varepsilon(\gamma_2)^{1/2}\| \\ &\quad + \|X_\infty^\varepsilon(\gamma_2)^{1/2}Y_\infty^\varepsilon(\gamma_2)X_\infty^\varepsilon(\gamma_2)^{1/2}\| \\ &\leq \|X_\infty^{1/2}(\gamma_2)Y_\infty(\gamma_2)X_\infty^{1/2}(\gamma_2) - X_\infty^\varepsilon(\gamma_2)^{1/2}Y_\infty^\varepsilon(\gamma_2)X_\infty^\varepsilon(\gamma_2)^{1/2}\| \\ &\quad + \|X_\infty^\varepsilon(\gamma_1)^{1/2}Y_\infty^\varepsilon(\gamma_1)(\gamma_1)^{1/2} - X_\infty^\varepsilon(\gamma_1)Y_\infty(\gamma_1)X_\infty^{1/2}(\gamma_1)\| \\ &\quad + \|X_\infty^{1/2}(\gamma_1)Y_\infty(\gamma_1)X_\infty^{1/2}(\gamma_1)\|. \end{aligned}$$

Since the obvious terms in the inequality chain tend to zero as $\varepsilon \rightarrow 0$, it follows that $\rho(X_\infty(\gamma_2)Y_\infty(\gamma_2)) \leq \rho(X_\infty(\gamma_1)Y_\infty(\gamma_1))$ for $0 < \hat{\gamma} \leq \gamma_1 \leq \gamma_2 \leq \infty$. \square

5. Rank. By Theorem 3.5, the rank of $X(\gamma)$ is a decreasing function of γ , because $X_+(\infty) \leq X(\gamma)$ for $0 < \gamma \leq \infty$. In fact, a more complete description of the rank is possible; namely, the rank of $X(\gamma)$ is a constant for all γ and it is determined by the eigenvalues of the dynamical system matrix A . According to Postlethwaite, Gu, and Young [20] the rank of $X_+(\infty)$ is determined by the unstable controllable eigenvalues and by the stable observable eigenvalues. Their descriptive and interesting eigenvalue characterization of the rank of $X_+(\infty)$ is also valid for $X(\gamma)$. The strategy of determining the rank of $X(\gamma)$ by using the direct sum decomposition of R^n derived in Lemma 3.1 is apparently new.

THEOREM 5.1. *If $X(\gamma)$ is a nonzero stabilizing solution of $\text{Ric}(X, H_\infty(\gamma)) = 0$ for some γ , $0 < \gamma \leq \infty$, then the rank of $X(\gamma)$ equals the number of unstable eigenvalues of A that are $(A, (1/\gamma^2)B_1B_1^T - B_2B_2^T)$ -controllable plus the number of stable eigenvalues of A that are (C, A) -observable. Moreover, if $X(\gamma)$ is stabilizing on some γ -interval $0 < \gamma_1 \leq \gamma \leq \gamma_2 \leq \infty$ then rank of $X(\gamma)$ is constant on the interval.*

Proof. By assumption, the closed-loop matrix $A + ((1/\gamma^2)B_1B_1^T - B_2B_2^T)X(\gamma)$ is stable. The decomposition in Lemma 3.1 can be realized by a real orthogonal similarity transformation. Since the eigenvalues of the A_{22} block are feedback invariants, the transformed closed loop matrix will be stable. Now, the dimension of $X_{11}(\gamma)$ equals the dimension of the block matrix A_{11} . The unstable eigenvalues (if any) of A_{11} are clearly controllable. Since $X_{11}(\gamma)$ is nonsingular, (3.1) can be rearranged to yield the equivalent equation

$$(5.1) \quad \begin{aligned} A_{11} + \left(\frac{1}{\gamma^2} B_{11}B_{11}^T - B_{21}B_{21}^T \right) X_{11}(\gamma) \\ = -X_{11}^{-1}(\gamma)(A_{11} + X_{11}^{-1}(\gamma)C_{11}^TC_{11})^TX_{11}(\gamma) \end{aligned}$$

Equation (5.1) clearly shows that $(A_{11} + X_{11}^{-1}(\gamma)C_{11}^TC_{11})$ is antistable, and so the stable eigenvalues of A_{11} (if any) are clearly observable, while the decomposition in Lemma 3.1 shows that the stable eigenvalues of A_{22} are clearly (C, A) -unobservable. Since these two types of eigenvalues exhaust the possibilities, the rank of $X(\gamma)$ must be as asserted.

Suppose that the rank of the stabilizing solutions $X(\gamma)$ is not constant on some open interval. Since the number of unstable eigenvalues of A are independent of γ , the rank of $X(\gamma)$ can only change on the stable eigenvalues. If an eigenvalue λ of A_{11} is (C_{11}, A_{11}) -observable, then it is also (C, A) -observable, because all right eigenvectors of A_{22} are members of the subspace $\text{Ker}(C)$ and all right eigenvectors of A_{11} are members of the subspace $\text{Im}(C)$, and the dimensions $\text{Im}(C)$, A_{11} and $\text{Im} X(\gamma)$ are equal. If the rank $X(\gamma_1)$ is different from the rank of $X(\gamma_2)$ on the interval $\gamma_1 \leq \hat{\gamma}_1 < \hat{\gamma}_2 \leq \gamma_2$, then, by using Lemma 3.1 again, there will exist a stable eigenvalue of A that is both (C, A) -observable and (C, A) -unobservable, which is impossible. \square

Appendix. The main theorem in this appendix is an amalgamation of separate theorems—some stated with more generality—that have been proved in recent publications. The theorem itself has not appeared in any of the separate publications and it will not be proved here. For completeness, each of the contributing results are referenced and they can be consulted for the relevant proofs.

THEOREM A. *Suppose that M and Q are real symmetric matrices with $Q \geq 0$ and ε is a nonnegative parameter.*

The matrix Riccati equation

$$(A.1) \quad A^T P + PA - PMP + Q = 0$$

has a unique stabilizing solution $P \geq 0$ if and only if

(i) *the Hamiltonian*

$$H = \begin{pmatrix} A & -M \\ -Q & -A^T \end{pmatrix}$$

has no imaginary eigenvalues, and

(ii) *the matrix Riccati equation for*

$$(A.2) \quad A^T P^\varepsilon + P^\varepsilon A - P^\varepsilon M P^\varepsilon + Q + \varepsilon I = 0$$

has a singular unique stabilizing solution $P^\varepsilon > 0$ for ε ($\varepsilon > 0$) sufficiently small that depends continuously on ε . Moreover, if P is stabilizing for $\text{Ric}(P, H)$ then the stabilizing solution P^ε for (A.2) will exist.

This fundamental result shows that the stabilizing solution of (A.1) will retain its property in the proximity of a perturbed Riccati equation. The solution depends continuously—at least locally—on the coefficient variation in the Riccati equation. By using a norm-induced metric space and the implicit function theorem, Gahinet and Laub [5] formulate the previous sentence in precise mathematical concepts. By a clever use of the implicit function theorem, they show in a neighborhood—induced by the metric norm—of the stabilizing solution set (P, A, M, Q) the matrix Riccati equation for the perturbed matrices $(A + \Delta A, M + \Delta M, Q + \Delta Q)$ will also have a stabilizing solution, provided that ΔA , ΔM , and ΔQ are real matrices, $\Delta M^T = \Delta M$, $\Delta Q^T = \Delta Q$ and that they are “sufficiently close” to the normal solution set. This local result does not require any other special system requirements such as stabilizability, observability, or inertia properties for ΔM or ΔQ . Another and earlier proof of the regularity of the stabilizing solution of the matrix Riccati equation using the implicit function theorem is found in [3].

Petersen, Anderson, and Jonckheere [18] prove that the existence of a positive semidefinite symmetric solution of (A.2) implies that (A.1) will have a positive semidefinite strong solution. Actually, the matrix $Q + \varepsilon I$ in (A.2) can be replaced by $\hat{Q} \geq Q$ with $\hat{Q} > 0$ in their theorem.

When $\varepsilon > 0$ $(A, \varepsilon I)$ is clearly controllable, and so by a well-known result of Shayman [23], every real symmetric solution of (A.2) is nonsingular. Finally, the statement “ $P \geq 0$ then $P^\varepsilon > 0$ ” in the theorem is proved in paper by Safonov, Limebeer, and Chiang [24]. Whenever (A.1) has a stabilizing solution P , then (A.2) for ε sufficiently small will have a stabilizing solution P^ε also, and both of these solutions will be unique by a theorem in Petersen [19].

Acknowledgments. The author thanks the anonymous referees for their helpful comments.

REFERENCES

- [1] D. BERNSTEIN AND W. M. HADDAD, LQG control with an H_∞ performance bound: a Riccati equation approach, IEEE Trans. Automat. Control, AC-34 (1989), pp. 293–305.
- [2] S. P. BOYD, V. BALAKRISHNAN, AND P. KABAMBA, Bisection methods for computing the H^∞ -norm of a transfer matrix and related problems, Math. Control Signals Systems, 2 (1989), pp. 207–219.

- [3] P. F. DELCHAMPS, *A note on the analyticity of the Riccati metric*, Algebraic and Geometric Methods in Linear Systems Theory, Lecture Notes in Applied Mathematics, AMS, Providence, RI, 18 (1980), pp. 37–41.
- [4] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.
- [5] P. GAHINET AND A. J. LAUB, *Computable bounds for the sensitivity of the algebraic Riccati equation*, Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, Report No. SCL 89-10, May 1989. SIAM J. Control and Optim., 28 (1990), pp. 1461–1480.
- [6] K. GLOVER AND J. C. DOYLE, *State space formulae for all stabilizing controllers that satisfy an H^∞ bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.
- [7] L. GOHBERG, P. LANCASTER, AND L. RODMAN, *On Hermitian solutions of the symmetric algebraic Riccati equation*, SIAM J. Control Optim., 24 (1986), pp. 1323–1334.
- [8] G. HEWER AND C. S. KENNEY, *The sensitivity of the stable Lyapunov Equation*, SIAM J. Control Optim., 26 (1988), pp. 321–344.
- [9] D. HINRICHSSEN AND A. J. PRITCHARD, *A Riccati equation approach to maximizing the complex stability radius by state feedback*, Internat. J. Control, 52 (1990), pp. 769–794.
- [10] R. A. HORN AND C. A. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1984.
- [11] E. A. JONCKHEERE AND L. M. SILVERMAN, *A new set of invariants for linear systems-applications to reduced order compensator design*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 953–964.
- [12] C. S. KENNEY AND G. A. HEWER, *The sensitivity of the algebraic and differential Riccati equations*, SIAM J. Control Optim., 28 (1990), pp. 50–69.
- [13] V. KUCERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344–347.
- [14] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.
- [15] K. MARTENSSON, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17–49.
- [16] B. P. MOLINARI, *The time-invariant linear-quadratic optimal control problem*, Automatica, 13 (1977), pp. 347–357.
- [17] P. PANDEY, C. S. KENNEY, A. PACKARD, AND A. J. LAUB, *A gradient method for computing the optimal H_∞ norm*, in Proc. 29th IEEE Conference on Decision and Control, Honolulu, HI, 1990, pp. 2628–2629.
- [18] I. R. PETERSEN, B. D. O. ANDERSON, AND E. A. JONCKHEERE, *A first principles solution to the non-singular H^∞ control problem*, to appear.
- [19] I. R. PETERSEN, *Some new results on algebraic Riccati equations arising in linear quadratic differential games and the stabilization of uncertain linear systems*, Systems Control Lett., 10 (1988), pp. 341–348.
- [20] I. POSTLETHWAITE, D. W. GU, AND S. D. O. YOUNG, *Some computational results on size reduction in H^∞ design*, AC-33 (1988), pp. 177–185.
- [21] A. C. M. RAN AND A. VREUGDENHIL, *Existence and comparison theorems for algebraic Riccati equations for continuous and discrete-time systems*, Linear Algebra Appl., 99 (1988), pp. 63–83.
- [22] M. G. SAFONOV AND D. J. N. LIMEBEER, *Simplifying the H^∞ theory via loop shifting*, in Proc. 27th Conf. Decision and Control, Austin, TX, 1988, pp. 1399–1404.
- [23] M. A. SHAYMAN, *Geometry of the algebraic Riccati equation*, SIAM J. Control Optim., 21 (1983), pp. 375–409.
- [24] M. G. SAFONOV, D. J. N. LIMEBEER, AND R. X. CHIANG, *Simplifying the H^∞ theory via loop-shifting, matrix-pencil, and descriptor concepts*, Internat. J. Control, 50 (1989), pp. 2467–2488.
- [25] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971) pp. 621–634.
- [26] H. K. WIMMER, *Monotonicity of maximal solutions of algebraic Riccati equations*, Systems Control Lett., 5 (1985), pp. 317–319.
- [27] K. ZHOU AND P. P. KHARGONEKAR, *An algebraic Riccati equation approach to H^∞ optimization*, Systems Control Lett., 11 (1988), pp. 85–91.

A GLOBALLY CONVERGENT STOCHASTIC APPROXIMATION*

SID YAKOWITZ†

Abstract. By combining a constrained Kiefer–Wolfowitz search with an automatic learning algorithm, it is shown that asymptotically normal convergence of an estimator to a global optimum under reasonably lenient assumptions can be attained. It is enough that the objective function be smooth and locally strictly convex at its minima. The central conclusion is that if θ_n is the estimate produced by the method shown at the n th decision epoch, then for some global minimizer θ^* , $n^{1/3}(\theta_n - \theta^*)$ is asymptotically normally distributed. This coincides with the conventional Kiefer–Wolfowitz convergence rate to a local optimum.

Whereas this study was motivated by needs of machine learning, the basic plan would seem applicable to root-finding tasks, and to other types of stochastic approximation algorithms.

Key words. stochastic approximation, random search, automatic learning

AMS(MOS) subject classifications. primary 62G05; secondary 62G99

1. An algorithm and statement of results. Many problems in automatic learning can be abstracted as a stochastic minimization problem, to wit: Let $f(\cdot)$ be an unknown function defined on a given domain D . On the basis of noisy observation pairs $\{(\theta_i, Y_i)\}_{i=1}^{n-1}$, with

$$(1.1) \quad Y_i = f(\theta_i) + Z(\theta_i),$$

choose θ_n in such a fashion that with respect to some criterion,

$$(1.2) \quad f(\theta_n) \rightarrow f_{\text{MIN}}.$$

Here f_{MIN} designates the global minimum of $f(\cdot)$ on D , and $Z(\theta_i)$ is a 0-mean random variable depending on the past only through the choice of θ_i .

A number of investigators, e.g., Devroye [1], [2], Gurin [3], and Yakowitz and Lugosi [14], have provided stochastic minimization algorithms that are globally consistent in various senses, and Yakowitz and Fisher [11] and Yakowitz and Lowe [13] have established rates of convergence. These works did not assume the objective function to be smooth, as will be required here. However, under our more stringent assumptions, a much more rapid convergence rate is assured. Computer experimentation on sequences of games, puzzles, and model queueing processes (some of which is reported in Yakowitz [10] and Yakowitz and Lowe [13]) have convinced the author that automatic learning holds promise as a practical methodology. A feature of the objective functions $f(\cdot)$ from these areas of applications is that they are multimodal, and local minima preclude the use of conventional methods.

Neural networks have become a popular facet of learning theory. We may readily confirm (Khanna [5, Chap. 5]) that standard neural network algorithms resemble stochastic approximation formulas. For that reason, the present study may have relevance because a common complaint about neural network learning is that the algorithms are attracted to suboptimal local minima.

The present paper is devoted to synthesizing an automatic learning procedure with Kiefer–Wolfowitz-type stochastic approximations (see Kiefer and Wolfowitz [6]) so as to take advantage of the rapid convergence properties of the latter technique

* Received by the editors September 24, 1990; accepted for publication (in revised form) July 1, 1991. The author was partially supported by National Science Foundation grant ECS 89-13642 and National Institutes of Health grant RO1 AI29426.

† Systems and Industrial Engineering Department, University of Arizona, Tucson, Arizona 85721.

TABLE 1
A stochastic approximation learning algorithm.

The Components	
D , where D is convex, $D \subset \mathbb{R}^d$.	
$p(\cdot)$, a continuous probability density function (chosen by the user) with D as support.	
$\{N(i)\}$, with	
(1.3)	$N(i) = \text{integer part of } (C \exp(i)), \quad i = 1, 2, \dots$
In (1.3), C is an arbitrary positive constant. The $N(i)$'s are referred to as "new sample times."	
$\{a(i)\}$ and $\{c(i)\}$, are positive sequences, with	
	$a(i) = A/i \quad \text{and} \quad c(i) = C'/i^{1/6}.$
Here A and C' are positive constants, and the sequences are parameters of the K-W search.	
Initialization	
$n = 1, NP = 0$. (n is the decision time, NP is Number of test Points at time n .)	
The Procedure	
If $n \in \{N(i)\}$, then get new test point.	
1. Set $NP = NP + 1$.	
2. Choose a new test point $T_{NP}(n)$ at random according to the density $p(\cdot)$. Set $\theta_n = T_{NP}(n)$ and observe the noisy value	
(1.4)	$Y_n = f(\theta_n) + Z(\theta_n).$
3. Initialize a sample average, a resample counter, and a hypercube for the new test point by defining	
	$m_{NP}(n) = Y_n, \quad \text{and} \quad NS_{NP}(n) = 1,$
and define $H(NP)$ to be the hypercube centered at T_{NP} with sides of length $(1/NP)^{1/d}$. This hypercube is fixed for the rest of the process. (T_{NP} will be constrained to wander around inside of $H(NP)$ as the process evolves.)	
4. Select a good test point for resampling. If $NP > 1$, define MIN to be the index j that minimizes $m_j(n)$, $1 \leq j \leq NP(n)$. Designate as I^* the smallest index i such that	
(1.5)	$m_i \leq m_{\text{MIN}} + 2/\log(n).$
Go to 8.	
Else if $n \notin \{N(i)\}$ then take a K-W step at the (apparently) best point and update other points as needed.	
5. Make a K-W step at I^* : We describe the K-W for the one-dimensional domain case, but the extension to dimension d should be obvious. Set	
	$\theta_n = T_{I^*}$
	$NS = NS_{I^*}(n)$
	$a = a(NS)$
(1.6)	$c = c(NS)$
	$Y1 = f(\theta_n + c) + Z(\theta_n + c)$
	$Y2 = f(\theta_n - c) + Z(\theta_n - c)$
	$DY^*(\theta_n, c) = \frac{1}{2c} (Y1 - Y2).$
Set	
(1.7)	$T = T_{I^*} - a \times DY^*(\theta_n, c).$
The K-W step redefines T_{I^*} to be $T_{I^*} = T$ for $T \in H(I^*)$ or the closest point in $H(I^*)$ to T if $T \notin H(I^*)$.	
6. Update the sample mean and counter:	
(1.8)	$m_{I^*}(n) = 1/(NS + 1)[NSm_{I^*}(n-1) + (Y1 + Y2)/2]$
and	
	$NS = NS + 1.$
7. Update other test points if necessary: Do preceding step at any index i , $1 \leq i \leq NP$, such that $NS_i(n) < n^\gamma$. Here $\gamma \in (0.5, 1)$ is a number which is held fixed for the duration of the process.	
8. Set $n = n + 1$ and repeat the procedure.	

while preserving robust consistency characteristics of the former. It will suffice for our purposes that $f(\cdot)$ be smooth, at least piecewise, and locally strictly convex near its minima.

Table 1 gives a global Kiefer-Wolfowitz (K-W) rule based on the Yakowitz-Lugosi method. Basically, the (very conventional) idea is to occasionally explore the domain to find better points and to improve estimates of values at old but unpromising test points. However, at most decision times, we choose test points at which the performance, as measured by sample averages, is seemingly better. The purposes of this re-sampling are to (i) improve the estimate at these places and (ii) attain performance that is best with respect to the current state of knowledge.

The obvious but apparently new contribution we offer is to take K-W steps at these "resample" times. By this enhancement, we can hope that, as resampling proceeds, each test point "wanders" into the bottom of the valley in which it finds itself. By gradually acquiring new test points in the tradition of random search, eventually every valley will be explored.

The central result (§ 2) is that for the hybrid method, and θ^* some *global* minimizer of $f(\cdot)$,

$$(1.9) \quad n^{1/3}(\theta_n - \theta^*)$$

is asymptotically normal.

We close this section by presenting the globally convergent K-W algorithm and a simple illustration of its use.

An example. We close this introductory section by examining the experimental side. Figure 1 is a plot of

$$(1.10) \quad f(x) = x \sin(30x), \quad x \in D = [0, 2].$$

The noise is independent and identically distributed (i.i.d.) standard normal. Figure 2 gives the average observed performance $(1/n) \sum_{i=1}^n Y_i$ for a version of the K-W learning algorithm, and, for purposes of comparison, Fig. 3 was obtained from a code that is the same, except it does not take K-W steps, but keeps T_i constant at its original, randomly chosen value. That is, update (1.7) is omitted.

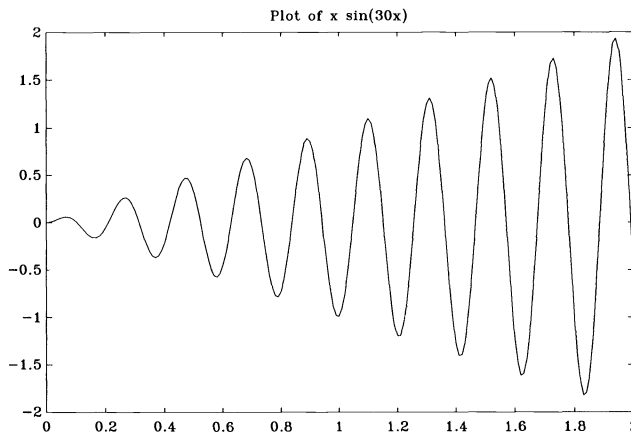


FIG. 1. The objective function.

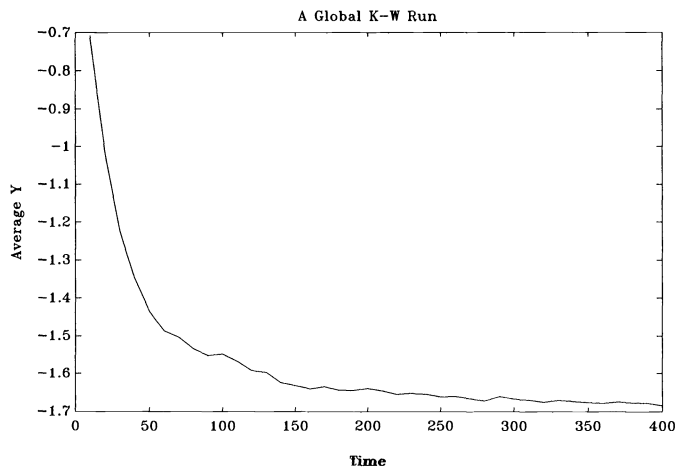


FIG. 2. A search with K-W steps, $\bar{Y}(n) = 1/n \sum_{1 \leq i \leq n} Y(\theta(i))$.

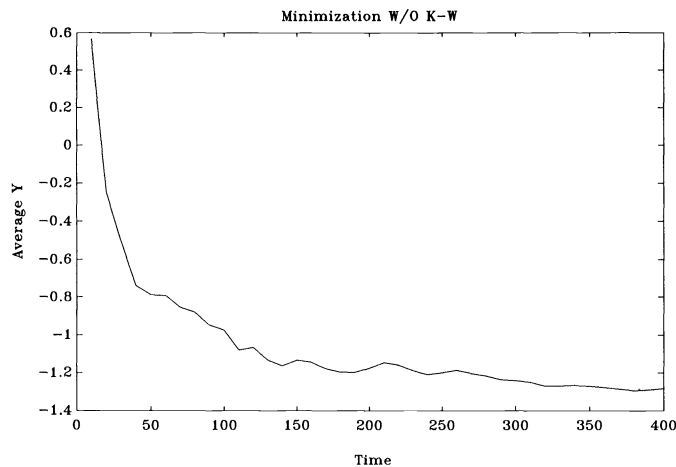


FIG. 3. A search without K-W steps.

Rationale and variations of the algorithm. The goal of finding a global stochastic approximation rule having been set, there are many plausible routes to its accomplishment. Alternative learning schemes that also could be synthesized with the K-W step are briefly discussed in Yakowitz and Lugosi [14]. The author suspects that the algorithm would be convergent without the hypercube constraint construct. However, a demonstration does not readily dawn on him. In any event, there is no hope of attaining a faster asymptotic rate than the present algorithm, which has the rate of the pure K-W process.

We anticipate that readers with specific stochastic minimization problems at hand will take liberties with our algorithm. To get an acceptable level of performance early in the learning process, it is sensible to initially take a goodly number of new point samples relatively soon. Perhaps computation in an interactive mode is warranted. If our global stochastic approximation methodology takes root, it may be worthwhile to devise an adaptive variation which chooses “new point” and “resample” times adaptively according to the degree of observed success in the exploration of new points,

and the observed sample variability in the running means $m_i(n)$. Such data-driven extensions are akin to “automatic bandwidth selection” procedures of the nonparametric estimation literature. To the extent that the nonparametric estimation literature is a useful guide, we may anticipate that analytic prowess will be needed to assure that convergence rates are maintained.

2. Convergence analysis. Here it is shown that a particular set of assumptions can assure convergence, in distribution and mean square, of a global stochastic approximation (SA) minimizer at a stated asymptotic rate. However, the reader will see from the plan we follow that other standard rules and assumptions of the SA realm can likewise be merged with the learning algorithm to attain convergence, perhaps in other senses. The first part of these developments pertains to pure K-W steps, and to avoid confusion with the learning algorithm, the decision variable is x , in §§ 2.1–2.3. Thus we seek a minimizer of $f(x)$ on the basis of noisy observations $Y(x) = f(x) + Z(x)$, $x \in D$. In § 2.4, we synthesize the constrained SA results with the global learning algorithm to demonstrate asymptotic convergence to a global minimizer.

2.1. Process assumptions.

About the objective function.

F.1. The function $f(\cdot)$ to be minimized is defined and three times continuously differentiable on a convex Borel set $D \subset R^d$. The domain D is further presumed to have an open subset.

F.2. The set of global minima (minimizers) of $f(\cdot)$ is finite but not empty.

F.3. Let f_{LOC} denote the infimum of the local but not global minima of $f(\cdot)$. Then

$$f_{\text{LOC}} > f_{\text{MIN}}.$$

F.4. Assume $f(\cdot)$ is locally strictly convex at its global minima, and each such minimum is an interior point of D . Furthermore, hypothesize that not all third-order partial derivatives are 0 at the minima.

About the noise.

N.1. The random variables $Z(x)$ are indexed by $x \in D$. Presume that $\{X(n)\}$ is a sequence obtained by some deterministic operation on observations

$$(X(i), Y(X(i))), i < n.$$

The distribution of $Z(X(n))$ depends on the past history only through the value $X(n)$. That is, in distribution, the conditioned variable

$$(2.1) \quad Z(X(n)) | \{X(1), Y(X(1)), \dots, X(n)\} = Z(X(n)) | X(n).$$

N.2. For each $x \in D$, $E[Z(x)] = 0$.

We remark that a consequence of (2.1) and the 0-conditional expectation assumption is that $\{Z(X(n))\}$ is a martingale difference sequence with respect to the sigma field induced by the preceding X - Y values.

N.3. For some positive constant σ^2 and all $x \in D$, the conditional variance $\text{var}(Z(x)) \leq \sigma^2$. The variance $\text{var}(Z(x))$ is further hypothesized to be continuous in neighborhoods of minima of $f(\cdot)$.

N.4. For some positive number δ and all x , we have

$$(2.2) \quad E|Z(x)|^{2+\delta} \leq M < \infty.$$

About the Kiefer–Wolfowitz steps.

S.1. The sequences $\{a(n)\}$ and $\{c(n)\}$ are defined by

$$(2.3) \quad a(n) = A/(n+1), \quad c(n) = C/(n+1)^{1/6},$$

A and C being positive constants.

S.2. Fix some d -dimensional hypercube $H \subset D$, and suppose the initial value $x(1)$ is chosen arbitrarily from H . For any real function (random or otherwise) $q(\cdot)$ and number $c \neq 0$, we use the notation

$$(2.4) \quad Dq(x, c) = (q(x+c) - q(x-c))/2c.$$

If $q(\cdot)$ is defined on a higher-dimension space, then $Dq(x, c)$ is a vector of like dimension. The j th coordinate is given by

$$Dq_j(x, c) = (q(x + ce_j) - q(x - ce_j))/2c,$$

where e_j is the vector with 1 at the j th coordinate and 0's elsewhere. With these constructs, the constrained search proceeds as follows: For $n \geq 1$, recursively define

$$(2.5) \quad \tilde{X}(n+1) = X(n) - a(n)DY(X(n), c(n)).$$

If $\tilde{X}(n+1) \in H$ then $X(n+1) = \tilde{X}(n+1)$. Otherwise, $X(n+1)$ is the closest vector in H to $\tilde{X}(n+1)$.

2.2. Almost sure convergence. The theory for almost sure convergence for constrained stochastic approximation and asymptotic normality offered in Kushner and Clark [7, Chaps. 5 and 7, respectively] is the foundation for our developments. Hereafter, “KC” will denote this reference and precede its equation and theorem numbers. This section and the following are devoted to recasting KC results into our setting. Our approach to demonstrating asymptotic normality requires first obtaining almost sure convergence, and that is the topic of discourse now.

LEMMA 2.1. *Let $f(\cdot)$ satisfy the conditions F.1, F.2, the noise process satisfy N.1–N.3, and the search be as in S.1 and S.2. Presume that $f(\cdot)$ assumes a minimum at x^* at an interior point of hypercube H and is strictly convex. Then almost surely,*

$$(2.6) \quad X(n) \rightarrow x^*.$$

Furthermore, if the convexity assumption is dropped, but $X(i)$ visits every neighborhood of x^ infinitely often, then almost surely, (2.6) holds.*

Proof. We proceed by showing that the assumptions of Theorem 5.3.1 [KC, p. 191] are satisfied, and then state how the conclusion implies the lemma. To apply this theorem, which ostensibly is directed at the Robbins–Monro case, we set $h(x) = -\nabla f(x)$, as suggested by [KC, p. 190]. Then β_n , as in [KC, eq. 5.3.1], is

$$(2.7) \quad \beta_n = -h(X(n)) - Df(X(n), c(n))$$

and

$$(2.8) \quad \xi_n = -DZ(X(n), c(n)).$$

From standard results on numerical differentiation.

$$(2.9) \quad \beta_n = O(c(n)^2).$$

Thus condition [KC, A5.1.5] is satisfied.

The condition [KC, A5.3.1] that the constraint functions be continuously differentiable is satisfied by the hyperplanes forming the boundary of H . Condition [KC, A5.3.2]

is satisfied if the series

$$(2.10) \quad \sum_{i=1}^{\infty} (a(i)/c(i))Z(X(i))$$

is almost surely convergent. But by N.3 and S.1, we have

$$\sum [a(i)/c(i)]^2 \sigma^2 < \infty$$

which for such orthogonal series, is known to imply almost sure convergence of (2.10) (e.g., Loève [8]).

We have now established that all the conditions of Theorem KC 5.3.1 are satisfied. The conclusion of the theorem is that if the point x is the limit of any convergent subsequence of $\{X(n)\}$, then x is a Kuhn-Tucker (KT) point. By the strict convexity assumption of our lemma, x^* is the unique KT point, and since H is compact, there must be a convergent subsequence. Thus (2.6) must hold. If convexity is dropped, the second part of the lemma is assured by the second part of Theorem KC 5.3.1.

2.3. Asymptotic normality. Theorem KC 7.3.1 addresses asymptotic normality of the unconstrained K-W algorithm directly. Under the conditions of Lemma 2.1, we have almost sure convergence to the minimizer, which is an interior point, and so eventually the points $X(n)$ are all determined by the pure K-W formula (2.5), and thus results about the unconstrained case apply.

LEMMA 2.2. *Under the conditions of Lemma 2.1, including the strict convexity assumption, and also N.4,*

$$(2.11) \quad n^{1/3}(X(n) - x^*)$$

converges in distribution to a normal vector with 0 mean. Also, the variable in (2.11) converges in mean square.

Proof. In the multivariable case, by “square” we mean right multiplication of the $d \times 1$ vector by its transpose. We follow the developments of [KC, Case 1, Chap. 7]. The result holds if we can show that the conditions of [KC, Thm. 7.3.1] are satisfied. Our choice of the sequences $a(n)$ and $c(n)$ are in agreement with [KC, A7.2.4(a)]. Lemma 2.1 is condition [KC, A7.2.2], Condition F.1 is [KC, A7.2.3]. Our noise assumptions N.1 to N.3 are [KC, A7.2.5] to [KC, A7.2.7]. This completes the hypotheses, and the conclusion of Theorem KC 7.3.1 is the conclusion of the lemma. (The discussion in [KC, § 7.4] makes this connection transparent.)

An expression for the limiting covariance matrix is given in [KC, 7.4.2], but it involves terms and parameters which will not usually be available to the statistician.

2.4. Convergence of the learning algorithm. In what follows, we refer to the times at which Algorithm Step 5 is taken, at which we resample at I^* , as *resample* times. The main result of our study is the following theorem.

THEOREM 2.1. *If all the assumptions F.1 to F.4, N.1 to N.4, and S.1 are in force, and the search algorithm of Table 1 is applied, then during resample times n , for some minimizer θ^* of $f(\cdot)$,*

$$(2.12) \quad n^{1/3}(\theta_n - \theta^*)$$

converges asymptotically in distribution to a normal random vector with zero mean. Moreover, the variable (2.12) converges in mean square to a constant matrix, and θ_n converges almost surely to θ^ .*

Proof. The plan is the following: We let V (for “valleys”) designate the union of neighborhoods of the global minima on which $f(\cdot)$ is strictly convex, and on which $f(\cdot)$ takes values no greater than $f_{\text{MIN}} + \varepsilon/2$, with $\varepsilon = f_{\text{LOC}} - f_{\text{MIN}}$ as in F.3. We will

show that eventually there will be hypercubes, as in Step 2, which lie entirely in V and which contain the minima as interior points. Almost surely, for all but finitely many n , resampling will concentrate on but one of these convergent test-points. It is then a simple matter to call upon the lemmas to confirm the statements of the theorem.

Let θ^* be a global minimum of $f(\cdot)$ and B_1 a ball centered at θ^* and having radius r sufficiently small that $f(\cdot)$ is convex on B_1 and is bounded above by $f_{\text{MIN}} + \varepsilon/2$ on B_1 . To show that eventually there is some hypercube $H(k)$ containing θ^* , and such that $H(k) \subset B_1$, take $N > 2/r$, and let E be the event that for some $i \geq N$, $\theta^* \in H(i)$. If E occurs, we have the desired hypercube. Let E^c be the complement of E . Then

$$\begin{aligned} P[E^c] &= \lim_n \prod_{i=N}^n (1 - P[\theta^* \in H(i)]) \\ (2.13) \quad &= \lim_n \prod_{i=N}^n (1 - [p(\theta^*)((1/i) + o(1/i))]) \end{aligned}$$

and this above product converges to 0.

Now that we know that almost surely any test point will eventually fall within an arbitrarily small search hypercube, the next goal is to show that for all but finitely many resample times, test point index I^* , as in Algorithm Step 4, will be selected to be some fixed index of a K-W search in V .

Let $\bar{f}_i(n)$ denote the average of all the true function values made at index i , $1 \leq i \leq NP(n)$, up to decision time n . That is, letting $\tau_i(j)$ denote the time of the j th call to test point T_i ,

$$(2.14) \quad \bar{f}_i(n) = 1/NS_i(n) \sum_{j=1}^{NS_i(n)} (f(\theta_{\tau_i(j)} + c(j)) + f(\theta_{\tau_i(j)} - c(j)))/2.$$

The actual averaged observations at index i , is, of course, the quantity $m_i(n)$, in the algorithm. The observation error due to noise is

$$\begin{aligned} (2.15) \quad e_i(n) &= \bar{f}_i(n) - m_i(n) \\ &= \sum_{j=1}^{NS_i(n)} (1/NS_i(n)) [Z(\theta_{\tau_i(j)} + c(j)) + Z(\theta_{\tau_i(j)} - c(j))]/2. \end{aligned}$$

The objective now is to demonstrate that

$$(2.16) \quad P[|e_i(N(j))| > 1/\log(N(j)), \text{ for infinitely many } i, j] = 0.$$

Fix an index i and let $S_i(n) = NS_i(n) \cdot e_i(n)$. Note that since Kolmogorov's inequality holds for martingales [4, p. 14], for any positive C ,

$$(2.17) \quad P\left[\max_{k \leq N(j)} |S_i(k)| > C\right] \leq N(j)\sigma^2/C^2.$$

From Step 7,

$$NS_i(N(j)) \geq N(j)^\gamma.$$

From this, and after setting $C = N(j)^\gamma/\log(N(j))$, and recalling that $NP(n) = O(\log(n))$, conclude that for some constant C' ,

$$\begin{aligned} (2.18) \quad &P\left[\max_{i \leq NP(N(j))} \max_{N(j)^\gamma \leq k \leq N(j)} |e_i(k)| > 1/\log(N(j))\right] \\ &\leq C' \log^3(N(j))\sigma^2/N(j)^{2\gamma-1}. \end{aligned}$$

The reader will verify that in view of the exponential growth rate of $N(j)$ (recall (1.3)), the right side of (2.18) is summable in index j . The Borel–Cantelli lemma then gives us (2.16).

Since we are examining asymptotic behavior, without loss of generality and in view of (2.16), in discussions to follow we assume that for all new sample times $N(j)$ at which I^* is reassigned,

$$(2.19) \quad |m_i(N(j)) - \bar{f}_i(N(j))| < 1/\log(N(j)), \quad 1 \leq i \leq NP(n).$$

We have noted that every global minimizer is eventually an interior point of some Step 3 hypercube on which the conditions of Lemma 2.1 are satisfied. The conclusion of Lemma 2.1 implies that every global minimizer is the target of some (in fact, arbitrarily many!) K-W searches $T_i(n)$. Define \mathcal{G} to be the (random) set of indices i such that $T_i(n) \rightarrow \bar{\theta}$, for some $\bar{\theta}$ a global minimizer of $f(\cdot)$, and designate G to be the minimum index in \mathcal{G} . By the second part of the Lemma 2.1 and our hypothesis F.3 (which implies that every global minimum is an isolated Kuhn–Tucker point), we conclude that for $i < G$ almost surely $\liminf \bar{f}_i(n) > \bar{f}_G(n)$. Consequently, in view of (2.16), only finitely many times can such a value $i < G$ serve as I^* . That is, almost surely, for all retest times n sufficiently large,

$$I^*(n) \geq G.$$

Toward analyzing the convergence behavior of $\bar{f}_G(n)$, we appeal to a law of the iterated logarithm for the K-W process in Hall and Heyde [4, § 7.6]. Hall and Heyde’s (designated now as HH) conditions B1 through B4 are covered by our smoothness and convexity postulates F1 and F4, and the observation that eventually $T_G(n)$ is a pure K-W process converging to θ^* . The HH condition B5 is satisfied by our noise assumptions N1 through N4. The HH analysis assumes that the K-W domain is of dimension 1, but assuming dimension $d > 1$ has influence only in replacing scalars by vectors, and reading absolute as norms, in their analysis. An implication of HH Theorem 7.15 is that almost surely,

$$(2.20) \quad |T_G(\tau_G(j)) - \theta^*| < C\sqrt{\log(\log(\tau_G(j)))/j^\gamma}$$

for some positive constant C and all j sufficiently large. After summing j from 1 to n^γ , and using that $f(T_G(\tau_G(j))) - f_{\min} = o(T_G(\tau_G(j)) - \theta^*)$, we see that almost surely,

$$(2.21) \quad \bar{f}_G(n) - f_{\min} = o(1/n^{1/6}).$$

Relations (2.16) and (2.21) yield that for all $i > G$ and large j ,

$$(2.22) \quad \begin{aligned} m_i(N(j)) + 2/\log(N(j)) &> f_{\min} + 2/\log(N(j)) \\ &> \bar{f}_G(n) + 2/\log(N(j)) + o(1/N(j)^{1/6}) \\ &> m_G(N(j)). \end{aligned}$$

The preceding sequence implies that condition (1.6) for choosing I^* will be satisfied by G for all but finitely many $N(j)$. This gives that almost surely, at all retest times n sufficiently large, Lemmas 2.1 and 2.2 apply to the search $T_G(n)$. The conclusions of these lemmas imply the theorem. \square

Upon recognizing that Step 1 and Step 7 times grow as $o(n)$, we can readily modify the global SA to assure asymptotic normality without restriction on n by letting the $N(i)$ ’s be randomly chosen (independently of the search process) integers from the interval

$$[C(1 - \alpha) \exp(i), C(1 + \alpha) \exp(i)],$$

for α an arbitrary number in the open unit interval and C as before. It is evident that the asymptotic normality result and convergence rate will be unchanged, but, of course, the almost sure convergence is sacrificed. The modifications of the proof of the theorem to cover this alteration are fairly evident.

COROLLARY 2.1. *If the conditions and the global SA are as in the theorem, except that the $N(i)$'s are randomly chosen from the integers in*

$$[C(1 - \alpha) \exp(i), C(1 + \alpha) \exp(i)],$$

then we have that, regardless of n ,

$$(2.23) \quad n^{1/3}(\theta_n - \theta^*)$$

converges asymptotically in distribution to a normal random vector with zero mean. Moreover, if the search domain D is bounded, the variable (2.23) converges in mean square to a constant matrix.

3. Conclusions. To some, the orientation of this study, with its emphasis on the "stochastic minimization" problem, will appear misguided. The major contribution would seem to be to the stochastic approximation method, about which we find much more research interest than machine learning. The orientation here stems from the author's conviction that the machine learning problem is the more significant, and that stochastic approximation is but one important tool among many for machine learning.

On the pragmatic side, the author and his students have found the global stochastic approximation to be effective for artificial-intelligence and heuristic search problems with low-dimension variables. Our experience with problems in which the search domain D has dimension higher than 10, say, has not been encouraging. This appears to have more to do with the curse of dimensionality than our deviation from pure K-W rules. (This phenomenon casts a shadow on the neural network enterprise.)

At the beginning of § 2, it was noted that our analysis could apparently be modified to encompass other stochastic approximation rules or to obtain convergence in other senses. In that regard, a fine study by Polak and Tsybakov [9] offers a stochastic approximation idea based on kernel regression notions and shows how to achieve optimal rates with respect to the class of all data-based nonparametric algorithms, under certain standard assumptions about the smoothness and noise.

As a final point, it is to be noted that our technique can be employed for the original Robbins-Monro task of finding the root of a regression function. Thus we could seek roots for functions $f(\cdot)$ which only satisfy the standard conditions locally, without sacrificing asymptotic rates.

Acknowledgments. This research effort stemmed from discussions with Professor Luc Devroye, who graciously invited me to a two-week visit at McGill University. Also, I am deeply indebted to a reviewer who not only uncovered a defect in the original proof of Theorem 2.1, but provided details of a way to circumvent the trouble and strengthen the conclusion, so as to show that G is almost surely chosen from some time onward. My gratitude is especially heartfelt because this reviewer offered these developments with utmost courtesy and diplomacy.

REFERENCES

- [1] L. P. DEVROYE, *The uniform convergence of nearest neighbor regression function estimators and their application in optimization*, IEEE Trans. Inform. Theory, 24 (1978), pp. 142-151.
- [2] ———, *Progressive global random search of continuous functions*, Math. Programming, 15 (1978), pp. 330-342.
- [3] L. S. GURIN, *Random search in the presence of noise*, Engrg. Cybernetics, 4 (1966), pp. 252-260.
- [4] P. HALL AND C. C. HEYDE, *Martingale Limit Theory and its Applications*, Academic Press, New York,

- [4] P. HALL AND C. C. HEYDE, *Martingale Limit Theory and its Applications*, Academic Press, New York, 1980.
- [5] T. KHANNA, *Foundations for Neural Networks*, Addison-Wesley, Reading, MA, 1990.
- [6] J. KIEFER AND J. WOLFOWITZ, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Stat., 23 (1952), pp. 462–466.
- [7] H. KUSHNER AND D. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1979.
- [8] M. LOÈVE, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, NJ, 1955.
- [9] B. T. POLYAK AND A. B. TSYBAKOV, *Optimal order of accuracy of search algorithms in stochastic optimization*, Problems Inform. Trans., 26 (1990), pp. 126–133.
- [10] S. YAKOWITZ, *A statistical foundation for machine learning with application to Go-moku*, Comput. Math. Appl., 17 (1989), pp. 1095–1102.
- [11] S. YAKOWITZ AND L. FISHER, *On sequential search for the maximum of an unknown function*, J. Math. Anal. Appl., 41 (1973), pp. 234–259.
- [12] S. YAKOWITZ AND M. KOLLIER, *Machine learning for blackjack counting strategies*, J. Statist. Plann. Inference, to appear.
- [13] S. YAKOWITZ AND W. LOWE, *Nonparametric bandits*, Ann. Oper. Res., 23 (1991), pp. 297–312.
- [14] S. YAKOWITZ AND E. LUGOSI, *Random search in the presence of noise, with application to machine learning*, SIAM J. Statist. Comput., 11 (1990), pp. 702–712.

OPTIMAL MUTUAL INFORMATION FOR CODERS AND JAMMERS IN MISMATCHED COMMUNICATION CHANNELS*

KENJIRO YANAGI†

Abstract. The mismatched communication channel with an infinite-dimensional real separable Hilbert space as input and output spaces is considered. To study communication in the presence of jamming, a two person zero sum game with mutual information of input source and output source as the payoff function can be formulated. The coder's goal is to make mutual information as large as possible, and the jammer's goal is to make mutual information as small as possible. The optimal mutual information under appropriate constraints of coders and jammers is obtained.

Key words. information theory, game theory, communications, jamming

AMS(MOS) subject classification. 94A

1. Introduction. The additive Gaussian channels can be considered in the following way. For the sake of simplicity, we consider both the input spaces and the output spaces to be a real separable Hilbert space H . Suppose that the noise source μ_Z is a Gaussian measure on H with mean 0 and covariance operator R_Z and the input source μ_X is a probability measure on H . Then the output source μ_Y is defined as

$$\mu_Y(A) = \mu_X \otimes \mu_Z\{(x, y); x + y \in A\}, \quad A \in \mathcal{B},$$

where $\mu_X \otimes \mu_Z$ is the usual product measure of μ_X and μ_Z and \mathcal{B} is the Borel σ -field of H . The compound source μ_{XY} derived from the input source μ_X and the noise source μ_Z is defined by

$$\mu_{XY}(B) = \mu_X \otimes \mu_Z\{(x, y); (x, x + y) \in B\}, \quad B \in \mathcal{B} \times \mathcal{B},$$

where $\mathcal{B} \times \mathcal{B}$ is the Borel σ -field of $H \times H$.

The mutual information $I(X, Y)$ of μ_{XY} with respect to $\mu_X \otimes \mu_Y$ is defined as follows: If $\mu_{XY} \ll \mu_X \otimes \mu_Y$,

$$I(X, Y) = \int_{H \times H} \log \frac{d\mu_{XY}}{d\mu_X \otimes \mu_Y}(x, y) d\mu_{XY}(x, y),$$

and otherwise $I(X, Y) = \infty$ (see [10], [11], [12], [17], [18]).

The information capacity is then $\sup \{I(X, Y); \mu_X \in \Phi\}$, where Φ is a set of admissible μ_X . Baker [1] defined mismatched Gaussian channels in the following way: Let μ_w be a Gaussian measure on H with mean 0 and covariance operator R_w satisfying

$$\text{range}(R_w^{1/2}) \subset \text{range}(R_Z^{1/2})$$

and

$$\overline{\text{range}(R_w)} = H.$$

Then there exists an unbounded densely defined selfadjoint operator S such that

$$R_Z = R_w^{1/2}(I + S)R_w^{1/2}.$$

* Received by the editors October 5, 1988; accepted for publication (in revised form) April 17, 1991.

† Department of Mathematics, Yamaguchi University, Yamaguchi, Japan.

An appropriate constraint is thus

$$\Phi = \left\{ \mu_X; \mu_X \text{ is a mean zero probability measure satisfying} \right. \\ \left. \int_H \|x\|_w^2 d\mu_X(x) \leq P(>0) \right\},$$

where $\|\cdot\|_w$ is the norm of reproducing kernel Hilbert space of μ_w . Then the capacity was obtained exactly.

McEliece and Stark [16] have modeled the conflict between coder and jammer when coding is used by a two-player zero-sum game with mutual information as the payoff function (see also [9], [19]). Recently, Hughes and Narayan [13], [14] obtained the interesting results for optimal coding. On the other hand, if the payoff function is instead taken as a quadratic distortion measure, it is known that Gaussian measures constitute saddle points for both finite and infinite-dimensional formulations (see [2], [3], [4], [5], [6], [7], [8]). But since we are interested in mutual information, we here adopt the viewpoint given in [9] that this is a game with two players. Player A, which we call the coder, controls the input source μ_X . Player B, which we call the jammer, controls the noise source μ_Z . The coder's goal is to make $I(X, Y)$ as large as possible, and the jammer's goal is to make it as small as possible. We call $I(X, Y)$ the game's payoff function in our game. This game will be meaningless and trivial unless we place restrictions on the players. We suppose that the coder's choice of μ_X must lie in a certain set Φ , the set of allowable inputs, and that μ_Z must lie in Ψ , the set of allowable noises. Then two programs are associated with this game.

$$\text{Coder's program:} \quad \alpha = \sup_{\mu_X \in \Phi} \inf_{\mu_Z \in \Psi} I(X, Y).$$

$$\text{Jammer's program:} \quad \beta = \inf_{\mu_Z \in \Psi} \sup_{\mu_X \in \Phi} I(X, Y).$$

A strategy μ_X^* such that

$$(1) \quad \inf_{\mu_Z \in \Psi} I(X^*, Y) = \alpha$$

is called an optimal strategy for the coder. The significance is that (1) implies

$$(2) \quad I(X^*, Y) \geq \alpha$$

for all allowable noises μ_Z . Hence, if the coder chooses the input μ_X^* , he is guaranteed a payoff of at least α , regardless of the jammer's strategy.

Similarly, an optimal strategy for the jammer is defined to be a strategy $\mu_Z^* \in \Psi$ such that

$$(3) \quad \sup_{\mu_X \in \Phi} I(X, Y^*) = \beta.$$

It follows that (3) implies

$$(4) \quad I(X, Y^*) \leq \beta$$

for all allowable input μ_X .

If it happens that $\alpha = \beta$, then combining (2) and (4), we have

$$(5) \quad I(X^*, Y^*) = \alpha = \beta$$

$$(6) \quad I(X, Y^*) \leq I(X^*, Y^*) \leq I(X^*, Y)$$

for every choice of allowable μ_X and μ_Z . If (5) holds, which is equivalent to (6), the common value is called the value of the game. The pair (μ_X^*, μ_Z^*) of optimal strategies is called a saddle point. In absence of other information, the coder will want to play strategy μ_X^* , and the jammer will want to play μ_Z^* .

2. Preliminaries. We assume that $\dim[H] = \infty$. We adopt the following Φ as a constraint of allowable coders:

$$\Phi = \left\{ \mu_X; \mu_X \text{ is a zero mean probability measure on } H \text{ satisfying} \right. \\ \left. \int_H \|x\|_W^2 d\mu_X(x) \leq P(>0) \right\}.$$

We also adopt the following Ψ as a constraint of allowable jammers:

$\Psi = \{\mu_Z; \mu_Z \text{ is a zero mean probability measure on } H \text{ with covariance operator } R_Z = R_W^{1/2}(I + S)R_W^{1/2}, \text{ where } S \text{ has } \theta \text{ (=the smallest limit point of the spectrum of } S) \text{ and } \{\lambda_n\}, \lambda_n \leq \lambda_{n+1} \text{ (=the set of eigenvalues of } S \text{ that are strictly less than } \theta), \text{ satisfying the following conditions:}$

$$\sum_n (\theta - \lambda_n) \geq P$$

and

$$\sum_n (1 + \lambda_n) \leq Q(>0)\}.$$

We remark that the limit points of the spectrum of S consist of all eigenvalues of infinite multiplicity, limit points of distinct eigenvalues, or points of the continuous spectrum. And we remark that $\#\{n; \lambda_n < \theta\} < \infty$, where $\#A$ denotes the number of elements of A . Then we set $L = \#\{n; \lambda_n < \theta\}$. The above Ψ in the infinite-dimensional channel is considered to be a naturally extended constraint of that which is stated in the finite-dimensional channel (Theorem 6). When μ_X is Gaussian with covariance operator

$$R_X = \sum_n \tau_n [R_Z^{1/2} u_n] \odot [R_Z^{1/2} u_n],$$

where $\tau_n \geq 0$ for $n \geq 1$, $\sum_n \tau_n < \infty$, $\{u_n; n \geq 1\}$ is a c.o.n. set and $(u \odot v)x = \langle x, v \rangle u$, and when μ_Z is also Gaussian, then we obtain

$$I(X, Y) = \frac{1}{2} \sum_n \log(1 + \tau_n).$$

Rewriting the condition of μ_X , we have

$$\sum_n \tau_n \|(I + S)^{1/2} U^* u_n\|^2 \leq P,$$

where U is a unitary operator. Setting $x_n^2 = \tau_n \|(I + S)^{1/2} U^* u_n\|^2$,

$$I(X, Y) = \frac{1}{2} \sum_n \log \left(1 + \frac{x_n^2}{1 + \lambda_n} \right)$$

where $\sum_n x_n^2 \leq P$, $\lambda_n = \langle S v_n, v_n \rangle$, $n \geq 1$, $\{v_n; n \geq 1\}$ is a c.o.n. set in the domain $\mathcal{D}(S)$ of S . We denote the direct sum by \oplus . Let R_n be the eigenspace of S relative to the eigenvalue λ_n . In order to make the value of game finite, we assume that

$$R_1 \oplus R_2 \oplus \cdots \oplus R_L \supset \text{linear supp}(\mu_X)$$

in Theorems 1-5 below, where $\text{linear supp}(\mu_X)$ means the linear support of μ_X .

When $\mu_Z \in \Psi$ is Gaussian, we at first obtain the following value:

$$C = \sup_{\mu_X \in \Phi} I(X, Y).$$

By the well-known results [1], [11], we obtain as follows:

1. If $\sum_{n=1}^L (\theta - \lambda_n) = P$, then

$$(7) \quad C = \frac{1}{2} \sum_{n=1}^L \log \frac{1+\theta}{1+\lambda_n} + \frac{1}{2} \frac{P + \sum_{n=1}^L (\lambda_n - \theta)}{1+\theta}.$$

2. If $\sum_{n=1}^L (\theta - \lambda_n) > P$, then the following hold:

(a) If $P + \sum_{n=1}^L \lambda_n > L\lambda_L$, then

$$(8) \quad C = \frac{1}{2} \sum_{n=1}^L \log \frac{\sum_{i=1}^L \lambda_i + P + L}{L(1+\lambda_n)}.$$

(b) If $K\lambda_{K+1} \geq P + \sum_{n=1}^K \lambda_n > K\lambda_K$ for some $K < L$,

$$(9) \quad C = \frac{1}{2} \sum_{n=1}^K \log \frac{\sum_{i=1}^K \lambda_i + P + K}{K(1+\lambda_n)}.$$

In both cases C is attained by Gaussian μ_X .

It is convenient to introduce the following notations:

$$\Phi^0 = \{\mu_X \in \Phi; \mu_X \text{ is Gaussian}\},$$

$$\Psi^0 = \{\mu_Z \in \Psi; \mu_Z \text{ is Gaussian}\},$$

$$\alpha_0 = \sup_{\mu_X \in \Phi^0} \inf_{\mu_Z \in \Psi^0} I(X, Y),$$

$$\beta_0 = \inf_{\mu_Z \in \Psi^0} \sup_{\mu_X \in \Phi^0} I(X, Y).$$

It is well known that we may show $\alpha_0 = \beta_0$ to show $\alpha = \beta$. And so, it is sufficient to show (6) for every $\mu_X \in \Phi^0$ and $\mu_Z \in \Psi^0$. That is, we can assume that both allowable coders and allowable jammers are Gaussian.

3. Statements and proofs. At first we state the following four theorems.

THEOREM 1. (L is fixed; θ and $\{\lambda_1, \lambda_2, \dots, \lambda_L\}$ are variable)

$$\alpha = \beta = \frac{L}{2} \log \left(1 + \frac{P}{Q} \right),$$

which is attained by Gaussian coder and Gaussian jammer satisfying $x_1^2 = x_2^2 = \dots = x_L^2 = P/L$, $\lambda_1 = \lambda_2 = \dots = \lambda_L = (Q - L)/L$, and $\theta > (P + Q)/L - 1$.

THEOREM 2. (L and θ are fixed; $\{\lambda_1, \lambda_2, \dots, \lambda_L\}$ are variable). Let $M = [(P + Q)/(1 + \theta)]$, when $[k]$ denotes the largest integer which is not larger than k .

1. If $L \leq M$, then

$$\alpha = \beta = \frac{L}{2} \log \frac{L(1+\theta)}{L(1+\theta) - P},$$

which is attained by Gaussian coder and Gaussian jammer satisfying $x_1^2 = x_2^2 = \dots = x_L^2 = P/L$, and $\lambda_1 = \dots = \lambda_L = \theta - P/L$.

2. If $L > M$, then

$$\alpha = \beta = \frac{L+1}{2} \log \left(1 + \frac{P}{Q} \right),$$

which is attained by Gaussian coder and Gaussian jammer satisfying $x_1^2 = x_2^2 = \dots = x_{L+1}^2 = P/(L+1)$, and $\lambda_1 = \lambda_2 = \dots = \lambda_{L+1} = Q/(L+1) - 1$.

THEOREM 3. (θ, L and $\{\lambda_1, \lambda_2, \dots, \lambda_L\}$ are all variable)

$$\alpha = \beta = \frac{1}{2} \log \left(1 + \frac{P}{Q} \right),$$

which is attained by Gaussian coder and Gaussian jammer satisfying

$$x_1 = P, \quad \lambda_1 = Q - 1, \quad \theta > P + Q - 1, \quad L = 1.$$

THEOREM 4. (θ is fixed; L and $\{\lambda_1, \lambda_2, \dots, \lambda_L\}$ are variable).

1. If $\theta > P + Q - 1$, then

$$\alpha = \beta = \frac{1}{2} \log \left(1 + \frac{P}{Q} \right),$$

which is attained by Gaussian coder and Gaussian jammer satisfying

$$x_1^2 = P, \quad \lambda_1 = Q - 1, \quad L = 1.$$

2. If $\theta \leq P + Q - 1$, then

$$\alpha = \beta = \frac{1}{2} \log \frac{\theta + 1}{\theta + 1 - P},$$

which is attained by Gaussian coder and Gaussian jammer satisfying

$$x_1^2 = P, \quad \lambda_1 = \theta - P, \quad L = 1.$$

Since the proofs of the above four theorems are similar, we only prove Theorem

2. Essentially, we need the following lemma.

LEMMA 1. The following inequalities hold.

1. If $x_i \geq 0 (1 \leq i \leq n)$, then

$$\log \left(1 + \frac{x_1}{a} \right) \left(1 + \frac{x_2}{a} \right) \dots \left(1 + \frac{x_n}{a} \right) \leq n \log \left(1 + \frac{x_1 + \dots + x_n}{na} \right).$$

The equality holds when $x_1 = x_2 = \dots = x_n$.

2. If $x_i > 0 (1 \leq i \leq n)$, then

$$n \log \left(1 + \frac{na}{x_1 + \dots + x_n} \right) \leq \log \left(1 + \frac{a}{x_1} \right) \left(1 + \frac{a}{x_2} \right) \dots \left(1 + \frac{a}{x_n} \right).$$

The equality holds when $x_1 = x_2 = \dots = x_n$.

Proof of Lemma 1.

1. Let $f(x) = 1 + x/a (x \geq 0)$ and $g(x) = \log f(x)$. Then $g''(x) = -1/(x+a)^2 < 0$. By Jensen's inequality,

$$g \left(\frac{x_1 + \dots + x_n}{n} \right) \geq \frac{1}{n} \{g(x_1) + \dots + g(x_n)\},$$

where the equality holds when $x_1 = x_2 = \dots = x_n$. Then we have

$$n \log \left(1 + \frac{x_1 + \dots + x_n}{na} \right) \geq \log \left(1 + \frac{x_1}{a} \right) \left(1 + \frac{x_2}{a} \right) \dots \left(1 + \frac{x_n}{a} \right).$$

2. Let $f(x) = 1 + a/x = (x+a)/x (x > 0)$ and $g(x) = \log f(x)$. Then $g''(x) = a(2x+a)/x^2(x+a)^2 > 0$. By Jensen's inequality,

$$g\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq \frac{1}{n} \{g(x_1) + \cdots + g(x_n)\},$$

where the equality holds when $x_1 = x_2 = \cdots = x_n$. Then we have

$$n \log \left(1 + \frac{na}{x_1 + \cdots + x_n}\right) \leq \log \left(1 + \frac{a}{x_1}\right) \left(1 + \frac{a}{x_2}\right) \cdots \left(1 + \frac{a}{x_n}\right). \quad \square$$

Proof of Theorem 2.

1. We show the left-hand side of (6). We have to maximize the value

$$\frac{1}{2} \sum_{n=1}^L \log \left(1 + \frac{x_n^2}{1 + \theta - P/L}\right),$$

where $\sum_{n=1}^L x_n^2 \leq P$, $x_n^2 \geq 0 (1 \leq n \leq L)$. By part 1 of Lemma 1,

$$\begin{aligned} & \log \left(1 + \frac{x_1^2}{1 + \theta - P/L}\right) \left(1 + \frac{x_2^2}{1 + \theta - P/L}\right) \cdots \left(1 + \frac{x_L^2}{1 + \theta - P/L}\right) \\ & \leq L \log \left(1 + \frac{x_1^2 + x_2^2 + \cdots + x_L^2}{L(1 + \theta) - P}\right). \end{aligned}$$

(The equality holds when $x_1^2 = \cdots = x_L^2$)

$$\begin{aligned} & \leq L \log \left(1 + \frac{P}{L(1 + \theta) - P}\right) \\ & = L \log \frac{L(1 + \theta)}{L(1 + \theta) - P}. \end{aligned}$$

(The equality holds when $x_1^2 + \cdots + x_L^2 = P$)

Then the maximal value is

$$\frac{2}{L} \log \frac{L(1 + \theta)}{L(1 + \theta) - P},$$

which is attained by $x_1^2 = \cdots = x_L^2 = P/L$.

Next we show the right-hand side of (6). We have to minimize the value

$$(10) \quad \frac{1}{2} \sum_{n=1}^L \log \left(1 + \frac{P/L}{1 + \lambda_n}\right),$$

where

$$\sum_{n=1}^L (1 + \lambda_n) \leq Q, \quad \sum_{n=1}^L (\theta - \lambda_n) \geq P.$$

Since $L \leq M$, we have $L \leq (P+Q)/(1+\theta)$. Then $L(1+\theta) - P \leq Q$. Hence we may minimize the value (10) where

$$\sum_{n=1}^L (\theta - \lambda_n) \geq P.$$

That is equal to $\sum_{n=1}^L (1 + \lambda_n) \leq L(1 + \theta) - P$. By part 2 of Lemma 1,

$$\begin{aligned} & \log \left(1 + \frac{P/L}{1 + \lambda_1} \right) \left(1 + \frac{P/L}{1 + \lambda_2} \right) \cdots \left(1 + \frac{P/L}{1 + \lambda_L} \right) \\ & \geq L \log \left(1 + \frac{P}{L + \lambda_1 + \lambda_2 + \cdots + \lambda_L} \right). \end{aligned}$$

(The equality holds when $\lambda_1 = \cdots = \lambda_L$)

$$\begin{aligned} & \geq L \log \left(1 + \frac{P}{L(1 + \theta) - P} \right) \\ & = L \log \frac{L(1 + \theta)}{L(1 + \theta) - P}. \end{aligned}$$

(The equality holds when $L + \lambda_1 + \cdots + \lambda_L = L(1 + \theta) - P$.)

Then the minimal value is

$$\frac{L}{2} \log \frac{L(1 + \theta)}{L(1 + \theta) - P},$$

which is attained by $\lambda_1 = \cdots = \lambda_L = \theta - (P/L)$.

2. We show the left-hand side of (6). We have to maximize the value

$$\frac{1}{2} \sum_{n=1}^{L+1} \log \left(1 + \frac{x_n^2}{Q/(L+1)} \right),$$

where $\sum_{n=1}^{L+1} x_n^2 \leq P$, $x_n^2 \geq 0$ ($1 \leq n \leq L+1$). By part 1 of Lemma 1 and the same method of Theorem 1, it is obtained that the maximal value is

$$\frac{L+1}{2} \log \left(1 + \frac{P}{Q} \right),$$

which is attained by $x_1^2 = \cdots = x_{L+1}^2 = P/(L+1)$.

Next we show the right-hand side of (6). We must maximize the value

$$(11) \quad \frac{1}{2} \sum_{n=1}^{L+1} \log \left(1 + \frac{P/(L+1)}{1 + \lambda_n} \right),$$

where

$$\sum_{n=1}^{L+1} (1 + \lambda_n) \leq Q, \quad \sum_{n=1}^{L+1} (\theta - \lambda_n) \geq P.$$

Since $L > M$, we have $(P + Q)/(1 + \theta) < L$. Then $(L+1)(1 + \theta) - P > Q$. By part 2 of Lemma 1 and the same method of Theorem 1, it is obtained that the minimal value is

$$\frac{L+1}{2} \log \left(1 + \frac{P}{Q} \right),$$

which is attained by $\lambda_1 = \cdots = \lambda_{L+1} = Q/(L+1) - 1$. \square

The following is an example in which the optimal coders and the optimal jammers do not exist.

THEOREM 5. (L and $\{\lambda_1, \lambda_2, \cdots, \lambda_L\}$ are fixed; θ is only variable); $\alpha_0 = \beta_0$ does not necessarily hold.

Counterexample of Theorem 5. Let $L = 2$, $\lambda_1 < \lambda_2$, $P > \lambda_2 - \lambda_1$, and $Q \geq 2 + \lambda_1 + \lambda_2$. The condition of θ is that $\theta \geq (P + \lambda_1 + \lambda_2)/2$. Then

$$\begin{aligned} e^{2\alpha_0} &= \max_{0 \leq x \leq P} \left[\min \left\{ \left(1 + \frac{x}{1 + \lambda_1}\right) \left(1 + \frac{P-x}{1 + \lambda_2}\right) \right. \right. \\ &\quad \left. \left. \cdot \left(1 + \frac{P-x}{1 + \lambda_1}\right) \left(1 + \frac{x}{1 + \lambda_2}\right) \right\} \right] \\ &= \left(1 + \frac{P}{2(1 + \lambda_1)}\right) \left(1 + \frac{P}{2(1 + \lambda_2)}\right). \end{aligned}$$

On the other hand, by the result of Baker [1] (see (7), (8), (9)),

$$e^{2\beta_0} = \frac{(P + 2 + \lambda_1 + \lambda_2)^2}{4(1 + \lambda_1)(1 + \lambda_2)}.$$

Hence $\beta_0 > \alpha_0$.

Finally we assume that $\dim[H] = N < \infty$. We adopt the following Φ as a constraint of allowable coders:

$\Phi = \{\mu_X; \mu_X \text{ is a zero mean probability measure on } H \text{ with covariance operator } R_X \text{ satisfying } \text{Tr}[R_X] \leq P(>0)\}$.

Since we can take $R_W = I$ (= the identity), $R_Z = R_W^{1/2}(I + S)R_W^{1/2}$ becomes $R_Z = I + S$. Then we can take $\theta = \infty$ because S has at most N eigenvalues. And $\{\lambda_n\}$ are all eigenvalues of S and $\{\lambda_n\}$ always satisfy the following conditions:

$$\sum_n (\theta - \lambda_n) \geq P.$$

So we use a true power constraint. Then we adopt the following Ψ as a constraint of allowable jammers:

$\Psi = \{\mu_Z; \mu_Z \text{ is a zero mean nondegenerate probability measure on } H \text{ with covariance operator } R_Z \text{ satisfying } \text{Tr}[R_Z] \leq Q(>0)\}$.

Now we obtain the final theorem.

THEOREM 6. *We have*

$$\alpha = \beta = \frac{N}{2} \log \left(1 + \frac{P}{Q}\right),$$

which is attained by Gaussian coder and Gaussian jammer satisfying $x_1^2 = x_2^2 = \dots = x_N^2 = P/N$ for eigenvalues of R_X and $r_1 = r_2 = \dots = r_N = Q/N$ for eigenvalues of R_Z .

Since we can prove this by the same method, we omit the proof.

4. Remarks.

1. To motivate the problem formulation, and to illustrate the main results, we give some specific examples of channels which would fit the infinite-dimensional Hilbert space set-up and satisfy the various assumptions made on the covariance operators associated with the input and noise signals.

The continuous time Gaussian channel is presented by

$$(12) \quad Y(t) = \int_0^t x(u) du + Z(t), \quad 0 \leq t \leq T,$$

where $x(\cdot)$, $Y(\cdot)$, and $Z(\cdot)$ are the channel input, the channel output, and the noise, respectively. The noise $Z(\cdot)$ is assumed to be a Gaussian process given by

$$Z(t) = B(t) + \int_0^t \int_0^s f(s, u) dB(u) ds,$$

where $B(\cdot)$ is a Brown motion and $f(s, u) \in L^2([0, T] \times [0, T])$ is a Volterra function (i.e., $f(s, u) = 0$ if $s < u$).

The white Gaussian channel is presented by

$$Y(t) = \int_0^t x(u) du + B(t), \quad 0 \leq t \leq T,$$

and is a special case of the Gaussian channel (12). We assume that an average power constraint

$$\int_0^T E[x(u)^2] du \leq PT$$

is imposed on the channel input, where $P > 0$ is a constant.

Let F and F^* be the integral operators on $L^2[0, T]$ with integral kernel $f(s, u)$ and $f^*(s, u) \equiv f(u, s)$, respectively, and define a selfadjoint operator S by

$$S = F + F^* + FF^*.$$

The S above works for jammers. Since S is a Hilbert-Schmidt operator, the smallest limit point of the spectrum of S is 0. Then we can assume that the negative eigenvalues $\{\lambda_n\}$, $\lambda_n \leq \lambda_{n+1}$, of S satisfy the following conditions:

$$\sum_n |\lambda_n| \geq P$$

and

$$\sum_n (1 + \lambda_n) \leq Q (> 0).$$

In Theorem 2, we assume that $\{\lambda_1, \dots, \lambda_L\}$ are variable and in Theorem 4, we assume that L and $\{\lambda_1, \dots, \lambda_L\}$ are variable.

2. We consider the problems that allow for the encoder to receive some feedback information from the output of the channel. In general, the following model for the discrete time Gaussian channel with feedback is considered:

$$Y_n = U_n + Z_n, \quad n = 1, 2, \dots, N,$$

where $Z = \{Z_n; n = 1, \dots, N\}$ is a nondegenerate, zero mean Gaussian process representing the noise and $U = \{U_n; n = 1, \dots, N\}$ and $Y = \{Y_n; n = 1, \dots, N\}$ are stochastic processes representing input signals and output signals, respectively. The channel is with noiseless feedback, so U_n is a function of a message X to be transmitted and the output signals Y_1, \dots, Y_{n-1} . We assume that a constraint, given in terms of the covariance matrix, is imposed on the input signals. Rigorously speaking, we assume the following constraints:

(A.1). A message X to be transmitted is a random variable, taking values in an arbitrary measurable space and independent of Z . However, we may regard messages $X = \{X_n; n = 1, \dots, N\}$ as stochastic processes.

(A.2). U_n is $\mathcal{F}(X) \vee \mathcal{F}_{n-1}(Y)$ -measurable, where $\mathcal{F}(X)$ and $\mathcal{F}_{n-1}(Y)$ are the σ -fields generated by X and $\{Y_k; k = 1, \dots, n-1\}$, respectively, and $\mathcal{A} \vee \mathcal{B}$ denotes the σ -field generated by σ -fields \mathcal{A} and \mathcal{B} .

(A.3). $\sum_{n=1}^N E[U_n^2] \leq P$.

Denote by Ω the class of all pairs (X, U) of a message X and an input signal U which satisfy the conditions (A.1)–(A.3). The mutual information quantity between a message X and the output signal $Y = \{Y_n; n = 1, \dots, N\}$ is denoted by $I(X, Y)$. Then the capacity C of the channel is defined as

$$C = \sup \{I(X, Y); (X, U) \in \Omega\}.$$

We denote $C_0(P)$ and $C_f(P)$ the capacities of Gaussian channels without and with feedback, respectively. In [15], [20], we obtained the necessary and sufficient conditions for the feedback capacity to increase.

Now we denote by $R_Z = \{z_{kl}\}$ the covariance matrix of Z with eigenvalues $0 < r_1 \leq r_2 \leq \dots \leq r_N$. We let $L_k = \{l(\neq k); z_{kl} \neq 0\}$. Then Z is said to be *white* when $L_k = \emptyset$ for any k , Z is said to be *blockwise white* when Z is not white and there exists k such that $L_k = \emptyset$, and Z is said to be *completely nonwhite* when $L_k \neq \emptyset$ for any k . When Z is blockwise white, we denote by \hat{R}_Z the submatrix of R_Z constructed by $\{k; L_k \neq \emptyset\}$.

The results are summarized in the following Proposition.

PROPOSITION 1. *The following results hold.*

- (a) *If Z is white, then $C_0(P) = C_f(P)$ for any P .*
- (b) *If Z is completely nonwhite, then $C_0(P) < C_f(P)$ for any P .*
- (c) *If Z is blockwise white, then:*
 - (i) *$P > P_0$ implies $C_0(P) < C_f(P)$,*
 - (ii) *$P \leq P_0$ implies $C_0(P) = C_f(P)$,*

where $P_0 = mr_m - (r_1 + \dots + r_m)$ and r_m is the smallest eigenvalues of \hat{R}_Z .

We apply our problems to the feedback case. We assume that $\dim[H] = N < \infty$. We denote the following Φ as a constraint of allowable coders:

$\Phi = \{(\mu_X, T); \mu_X \text{ is a zero mean probability measure on } H \text{ with covariance matrix } R_X, T \text{ is a Volterra matrix satisfying}$

$$\text{Tr}[(I + T)R_X(I + {}^tT) + TR_Z{}^tT] \leq P(>0)\},$$

where tT is the transposed matrix of T .

We adopt the following Ψ as a constraint of allowable jammers:

$$\Psi = \{\mu_Z; \mu_Z \text{ is a zero mean nondegenerate probability measure on } H \text{ with covariance matrix } R_Z \text{ satisfying } \text{Tr}[R_Z] \leq Q(>0)\}.$$

Using Proposition 1, it is not difficult to show the following result.

PROPOSITION 2.

$$\alpha = \beta = \frac{N}{2} \log \left(1 + \frac{P}{Q} \right),$$

which is attained by Gaussian coder and Gaussian jammer satisfying $x_1^2 = x_2^2 = \dots = x_N^2 = P/N$ for eigenvalues of R_X and $r_1 = r_2 = \dots = r_N = Q/N$ for eigenvalues of R_Z and $T = 0$.

Then we can see that feedback is not useful in our jamming channels.

Acknowledgments. The author would like to express his hearty thanks to the referee in the course of preparing this paper.

REFERENCES

- [1] C. R. BAKER, *Capacity of the mismatched Gaussian channel*, IEEE Trans. Inform. Theory, 33 (1987), pp. 802–812.
- [2] R. BANSAL AND T. BASAR, *Communication games with partially soft power constraints*, J. Optim. Theory Appl., 61 (1989), pp. 329–346.
- [3] T. BAŞAR, *A trace minimization problem with applications in joint estimation and control under nonclassical information*, J. Optim. Theory Appl., 31 (1980), pp. 343–359.
- [4] T. BAŞAR AND T. Ü. BAŞAR, *A bandwidth expanding scheme for communication channels with noiseless feedback in the presence of unknown jamming noise*, J. Franklin Inst., 317 (1984), pp. 73–88.

- [5] T. BAŞAR AND Y. W. WU, *A complete characterization of minimax and maximin encoder-decoder policies for communication channels with incomplete statistical description*, IEEE Trans. Inform. Theory, 31 (1985), pp. 482–489.
- [6] ———, *Solutions to a class of minimax decision problems arising in communication systems*, J. Optim. Theory Appl., 51 (1985), pp. 375–404.
- [7] T. Ü. BAŞAR AND T. BAŞAR, *Optimum coding and decoding schemes for the transmission of a stochastic process over a continuous-time stochastic channel with partially unknown statistics*, Statistics, 8 (1982), pp. 213–237.
- [8] ———, *Optimum linear causal coding schemes for Gaussian stochastic processes in the presence of correlated jamming*, IEEE Trans. Inform. Theory, 35 (1989), pp. 199–202.
- [9] J. M. BORDEN, D. M. MASON, AND R. J. McELIECE, *Some information theoretic saddlepoints*, SIAM J. Control Optim., 23 (1985), pp. 129–143.
- [10] R. L. DOBURUSHIN, *General formulation of Shannon's main theorem in information theory*, Uspekhi Mat. Nauk, 14 (1959), pp. 3–104.
- [11] R. G. GALLAGER, *Information Theory and Reliable Communication*, John Wiley, New York, 1968.
- [12] I. M. GEL'FAND AND A. M. YAGLOM, *Calculation of the amount of information about random functions contained in another such function*, Uspekhi Math. Nauk, 12 (1957), pp. 3–52.
- [13] B. HUGHES AND P. NARAYAN, *Gaussian arbitrarily varying channels*, IEEE Trans. Inform. Theory, (1987), pp. 267–284.
- [14] ———, *The capacity of a vector Gaussian arbitrarily varying channel*, IEEE Trans. Inform. Theory, 34 (1988), pp. 995–1003.
- [15] S. IHARA AND K. YANAGI, *Capacity of discrete time Gaussian channel with and without feedback II*, Japan J. Appl. Math., 6 (1989), pp. 245–258.
- [16] R. J. McELIECE AND W. E. STARK, *An information theoretic study of communication in the presence of jamming*, Proc. IEEE Internat. Conf. Communications, (1981), pp. 45.3.1–45.3.5.
- [17] M. P. PINSKER, *Information and information stability of random variables and processes*, Holden Day, San Francisco, 1964.
- [18] C. E. SHANNON, *A mathematical theory of communication*, Bell. Syst. Tech. J., 7 (1948), pp. 379–423.
- [19] W. E. STARK AND R. J. McELIECE, *On the capacity of channels with block memory*, IEEE Trans. Inform. Theory, 34 (1988), pp. 322–324.
- [20] K. YANAGI, *Necessary and sufficient condition for Gaussian feedback capacity to increase*, Proc. 1990 International Symposium on Information Theory and its Applications, November 27–30, Hawaii, 1990.

OPTIMAL CONTROL OF FAVORABLE GAMES WITH A TIME LIMIT*

MARTIN KULLDORFF†

Abstract. This paper shows how to optimally control a stochastic process if one seeks to reach a certain value before a fixed time without first hitting zero. The process has a drift in the favorable direction. Discrete time random walks, as well as continuous time diffusion processes, are considered. The controls available are such that zero variance means zero drift, and more variance means more drift. To be more precise, in the continuous time case it holds that $dX_t = \sigma(X_t, t)(\mu_t dt + dB_t)$, where σ is the control variable. As corollaries of the results, some interesting inequalities for stochastic processes are obtained.

Key words. stochastic control, gambling theory, red and black, limits of control problems, inequalities for stochastic processes, Brownian motion, random walk

AMS(MOS) subject classifications. 93E20, 60J15, 60J60

1. Introduction. The problem of red and black, which has taken its name from the game of roulette, has been of interest to probabilists for quite some time. It reads as follows. Suppose that we visit a casino with only one game available. In this game, we bet an amount s . With probability $1 - w$, we lose the stake, and, with probability w , we win the amount s . $w \in (0, 1)$ is fixed by the casino. The stake can be any amount of our choice, as long as it does not exceed the fortune f at that moment. Negative bets are not permitted. Our goal is, with repeated bets, to maximize the probability of reaching some fixed fortune $c > f$. Without loss of generality, we put $c = 1$, which implies that $f \in [0, 1]$. The question is how we should bet, and, with that probability, we reach the goal if we use the optimal strategy.

Although this problem has taken its structure from the world of casinos, the interesting applications are found in areas that are less glamorous. One example is the control of dams in connection to hydroelectric power plants [7]. The applications are not dealt with in this paper.

The problem might be divided into four cases: the odds could be favorable ($w > \frac{1}{2}$) for unfavorable ($w < \frac{1}{2}$), and the playing time could be finite or infinite.

Dubins and Savage [2] solved the unfavorable infinite time problem. They gave all optimal betting strategies as well as $U(f)$, the probability of reaching 1 if the initial fortune is f and if an optimal strategy is used. Bold play, i.e., betting the minimum of f and $1 - f$, is one of the optimal strategies.

The unfavorable finite time problem was solved by Dvoretzky (see [2, p. 92]). Also, bold play is one of the optimal strategies here.

In the favorable infinite time case, there are many strategies that reach the goal with probability one. They share the fact that the bets are timid, so that the law of large numbers ensures that the goal is eventually reached. Kelly [5] studied a related problem, where he sought to maximize the expected growth rate of the fortune. He gave the optimal betting strategy for this objective, which is betting a fixed proportion of the fortune every time. The more favorable the game is, the larger this proportion should be. This strategy of betting is now called the “Kelly criterion” and is also one of the optimal strategies for the favorable infinite time case.

* Received by the editors March 26, 1990; accepted for publication (in revised form) June 3, 1991.

† Department of Statistics, Uppsala University, P.O. Box 513, S-751 20 Uppsala, Sweden. Most of the research was done while the author was at the School of Operations Research and Industrial Engineering at Cornell University, Ithaca, New York 14853.

The fourth and last case remains, that of the favorable finite time problem. This is, together with some generalizations, its limits, and its continuous time version, the subject of this paper. Because of the favorability, bold play can no longer be expected to be optimal, and, because of the time limit, timid play cannot be optimal. The solution must lie somewhere in between. The problem was studied by Breiman [1]. He determined $U(f, t)$, the probability of reaching 1 with fortune f , and t times to play, if we play optimally. He also gave a method for finding one of the optimal strategies.

In practice, the Kelly criterion has often been used for these problems as well. The motivation for this has been based on intuition, a limit result of Breiman and the fact that nothing better has been available. Breiman [1] showed that $U(f, t) - K(f, t)$ converges to zero uniformly as $t \rightarrow \infty$. Here K is the probability of reaching 1 if we use the Kelly criterion. This result is not as strong as we would like. For $\varepsilon > 0$, consider the following strategy: First, we discard our entire fortune, except ε (if $f \leq \varepsilon$, discard nothing), and then we play according to the Kelly criterion. Let $K_\varepsilon(f, t)$ be the probability of reaching 1 if we use this strategy. It is clearly not a good strategy if ε is small, but $U(f, t) - K_\varepsilon(f, t)$ converges to zero uniformly for any $\varepsilon > 0$. To see this, note that $K(f, t) = K_\varepsilon(f, t)$ for $f \leq \varepsilon$ and that $K(\varepsilon, t) \rightarrow 1$ as $t \rightarrow \infty$.

In § 2.2 we give an alternative solution to that of Breiman. This gives us simple characteristics of $U(f, t)$ and of an entire family of optimal betting strategies, and it enables us to obtain their limits in § 2.6. Just as Dubins and Savage [2] extended their result to general primitive casinos, i.e., where we, instead of s , gain $((1-r)/r)s$, $r \in (0, 1)$ if we win, we do the same in § 2.4. In § 2.3 we replace the 0/1 utility of reaching or not reaching 1 with the utility function $u(f) = f$, $f \in [0, 1]$, where $u(f)$ is the value for us to have fortune f with no time left to play. Section 2.5 extends the results further to a family of utility functions.

More interesting than the discrete time problems, perhaps, are their continuous time counterparts. Heath et al. [4] solved the continuous time equivalent of Kelly's problem, and Sudderth and Weerasinghe [6] showed that bold play is optimal for unfavorable, continuous time, red and black when there is a time limit. In § 3 we solve the continuous time version of the favorable limited time problem, both for 0/1, $u(f) = f$ and some more general utility functions. The solutions prove to be the same as the limit of the discrete time problem.

Section 4 gives some inequalities for stochastic processes that follow as corollaries to the preceding results.

There is no dependence between §§ 2 and 3, so they can be read independently. The notational framework is that of Dubins and Savage [2].

2. Discrete time primitive casinos.

2.1. Description of problem. In the general discrete time problem, for a limited number of times, we place stakes $s(f, t)$. These might depend on f , the fortune, and t , the number of times left to play. Of course, $0 \leq s(f, t) \leq f$. With probability $1 - w$, we lose the stake, and, with probability w , we win the amount $((1-r)/r)s$. The numbers $w \in (0, 1)$ and $r \in (0, 1)$ are fixed to us. Since we are studying a favorable case, we have that $w > r$.

There is a utility function $u(f)$, describing the value of having fortune f when time is out. Our goal is to maximize the expected utility at the end of the game. We let $U(f, t)$ be the optimal expected utility at fortune f with t times left to play. Clearly, $U(f, 0) = u(f)$. We wish to find $U(f, t)$, as well as the strategy, i.e., sequence of stakes, that will give the optimal expected utility.

For readability, we use the following notation:

$$F(k|t, p) = \begin{cases} \sum_{i=0}^k \binom{t}{i} p^i (1-p)^{t-i} & \text{if } k \in \{0, 1, \dots, t\}, \\ 0 & \text{if } k < 0, \\ 1 & \text{if } k > t. \end{cases}$$

This is, of course, $P(X \leq k)$ when $X \sim \text{Binomial}(t, p)$.

DEFINITION. Let $q(f, t)$ and $k(f, t)$ be the unique numbers such that

$$f = F(k-1|t, 1-r) + q(f, t) \binom{t}{k} (1-r)^k r^{t-k} \quad \text{and} \quad 0 \leq q(f, t) < 1.$$

We often write k instead of $k(f, t)$. Understanding F , q , and k is important in the reading of the proofs.

2.2. Red and black. We first discuss the simplest case: red and black. We have that $u(f) = 0$ for $f < 1$, $u(1) = 1$, $r = \frac{1}{2}$, and $w > \frac{1}{2}$. The intuition we get from solving this problem enables us to do the general case in the next sections, as well as the limit problem. The proofs are only sketched or hinted at in this section, since these results follow from those in § 2.4. For an alternative approach to this problem, see Breiman [1]. He found U and gave a method for obtaining an optimal strategy.

DEFINITIONS. A fortune f is binary at time t , i.e., with t times left to play, if $f2^t$ is an integer.

A stake is binary at fortune f if we arrive at a binary fortune, regardless of whether we win or lose.

A strategy is binary if it only uses binary stakes.

By using a binary strategy, the player finishes at fortune 0 or 1. The pay-off from any optimal strategy has the following property.

PROPOSITION 1. *It holds that $U(f, t) = U(n/2^t, t)$, where n is the integer such that $n/2^t \leq f < (n+1)/2^t$.*

Proof. The proof follows by induction.

This means that $U(f, t)$ is a step function of f with jumps at the binary fortunes.

PROPOSITION 2. *If the initial fortune is binary, then every optimal strategy is binary.*

Proof. For the proof, use Proposition 1.

PROPOSITION 3. *If we play t times, there are 2^t possible outcomes of the gamble. If the initial fortune is binary with $f = n/2^t$ and the strategy is binary, then exactly n of the 2^t outcomes result in reaching 1, and the remaining $2^t - n$ outcomes result in reaching zero.*

Proof. We surely reach 0 or 1 if we use a binary strategy, and, if $w = \frac{1}{2}$, we get a martingale for the fortune, so $P(\text{reaching } 1) = U(n/2^t, t) = n/2^t$ for integers n . Since all specific outcomes have the same probability $1/2^t$, exactly n outcomes bring us to 1. Hence the proposition is true for $w = \frac{1}{2}$. Since the resulting fortune of an outcome does not depend on w , it is true for all w . \square

For $w > \frac{1}{2}$ we can get an upper bound on U by adding the n outcomes with the highest probabilities.

COROLLARY 1. *Let $k = k(n/2^t, t)$. If n is an integer, then*

$$U\left(\frac{n}{2^t}, t\right) \leq F(k-1|t, 1-w) + q\left(\frac{n}{2^t}, t\right) \binom{t}{k} (1-w)^k w^{t-k}.$$

The result is that this is not only an upper bound, but also the actual value of U . To show this, we must find a strategy that conserves U . In the following, including Proposition 4, note that we are not dealing with probabilities but only with outcomes (not being concerned about how likely they are), fortunes, and stakes.

With an initial fortune of $n/2^t$, we need all outcomes with at most $k(n/2^t, t) - 1 = k - 1$ losses to bring us to 1, and there can be no outcome with more than k losses that brings us there.

Suppose now that we win our first gamble. There are $F(k - 1 | t - 1, \frac{1}{2})2^{t-1}$ outcomes that start with a win and have at most $k - 1$ losses. Since all these must lead us to our goal, our new fortune in the case of a win must be at least $F(k - 1 | t - 1, \frac{1}{2})$. Likewise, the new fortune must be at most $F(k | t - 1, \frac{1}{2})$. So

$$F\left(k - 1 | t - 1, \frac{1}{2}\right) - \frac{n}{2^t} \leq s\left(\frac{n}{2^t}, t\right) \leq F\left(k | t - 1, \frac{1}{2}\right) - \frac{n}{2^t}.$$

If our first gamble is a loss instead, there are $F(k - 2 | t - 1, \frac{1}{2})2^{t-1}$ outcomes that start with a loss and have a total of at most $k - 1$ losses. This means that, in the case of a loss, we must finish with a fortune of at least $F(k - 2 | t - 1, \frac{1}{2})$. Likewise, our new fortune cannot exceed $F(k - 1 | t - 1, \frac{1}{2})$. So

$$\frac{n}{2^t} - F\left(k - 1 | t - 1, \frac{1}{2}\right) \leq s\left(\frac{n}{2^t}, t\right) \leq \frac{n}{2^t} - F\left(k - 2 | t - 1, \frac{1}{2}\right).$$

Putting these inequalities together, we get the following proposition.

PROPOSITION 4. *If f is a binary fortune with t times left to play, then $s(f, t)$ is an optimal stake if and only if*

- (i) $s(f, t)$ is binary,
- (ii) $s(f, t) \leq \min \{n/2^t - F(k - 2 | t - 1, \frac{1}{2}), F(k | t - 1, \frac{1}{2}) - n/2^t\}$,
- (iii) $s(f, t) \geq |F(k - 1 | t - 1, \frac{1}{2}) - n/2^t|$.

To verify that such a stake always exists, use the formula

$$F(k - 1 | t, \frac{1}{2}) = \frac{1}{2} F(k - 1 | t - 1, \frac{1}{2}) + \frac{1}{2} F(k - 2 | t - 1, \frac{1}{2})$$

and the fact that $F(k - 1 | t, \frac{1}{2}) \leq f \leq F(k | t, \frac{1}{2})$.

We also have the next proposition.

PROPOSITION 5 (see Breiman [1]). *It holds that*

$$U(f, t) = F(k - 1 | t, 1 - w) + q\left(\frac{n}{2^t}, t\right) \binom{t}{k} (1 - w)^k w^{t-k},$$

where n is the largest integer such that $f \geq n/2^t$.

2.3. Utility function $u(f) = f$. We now consider the general primitive casino, where r is not necessarily equal to $\frac{1}{2}$. In this section we will consider the case in which $u(f) = f$. We first need a simple combinatorial lemma.

LEMMA 1. *Let $k = k(f, t)$. If $F(k - 1 | t - 1, 1 - r) \leq f \leq F(k | t - 1, 1 - r)$ (i.e., $k/t \leq q(f, t) < 1$), then*

$$\begin{aligned} f &= F(k - 1 | t, 1 - r) + q(f, t) \binom{t}{k} (1 - r)^k r^{t-k} \\ &= F(k - 1 | t - 1, 1 - r) + r \frac{t}{t - k} \left(q(f, t) - \frac{k}{t} \right) \binom{t - 1}{k} (1 - r)^k r^{t-1-k}. \end{aligned}$$

If $F(k-1|t, 1-r) \leq f \leq F(k-1|t-1, 1-r)$ (i.e., $0 \leq q(f, t) < k/t$), then

$$\begin{aligned} f &= F(k-1|t, 1-r) + q(f, t) \binom{t}{k} (1-r)^k r^{t-k} \\ &= F(k-2|t-1, 1-r) + \left[r + \frac{t}{k} (1-r) q(f, t) \right] \binom{t-1}{k-1} (1-r)^{k-1} r^{t-k}. \end{aligned}$$

Proof. It holds that

$$\begin{aligned} &F(k-1|t, 1-r) + q(f, t) \binom{t}{k} (1-r)^k r^{t-k} \\ &= rF(k-1|t-1, 1-r) + (1-r)F(k-2|t-1, 1-r) \\ &\quad + q(f, t) \binom{t}{k} (1-r)^k r^{t-k} \\ &= F(k-2|t-1, 1-r) + r \binom{t-1}{k-1} (1-r)^{k-1} r^{t-k} \\ &\quad + \frac{t}{k} (1-r) q(f, t) \binom{t-1}{k-1} (1-r)^{k-1} r^{t-k} \\ &= F(k-2|t-1, 1-r) + \left[r + \frac{t}{k} (1-r) q(f, t) \right] \binom{t-1}{k-1} (1-r)^{k-1} r^{t-k}. \end{aligned}$$

If $q(f, t) \geq k/t$, then $r + (t/k)(1-r)q(f, t) \geq 1$; so the above is equal to

$$\begin{aligned} &F(k-2|t-1, 1-r) + \binom{t-1}{k-1} (1-r)^{k-1} r^{t-k} + \left[\frac{t}{k} (1-r) q(f, t) + r - 1 \right] \binom{t-1}{k-1} \\ &\quad \cdot (1-r)^{k-1} r^{t-k} \\ &= F(k-1|t-1, 1-r) + (1-r) \left[\frac{t}{k} q(f, t) - 1 \right] r \frac{k}{t-k} \binom{t-1}{k} (1-r)^{k-1} r^{t-1-k} \\ &= F(k-1|t-1, 1-r) + r \frac{t}{t-k} \left[q(f, t) - \frac{k}{t} \right] \binom{t-1}{k} (1-r)^k r^{t-1-k}. \quad \square \end{aligned}$$

THEOREM 1. *Let*

$$\begin{aligned} Q(f, t) &= F(k-1|t, 1-w) + q(f, t) \binom{t}{k} (1-w)^k w^{t-k} \quad \text{and} \\ f &= F(k-1|t, 1-r) + q(f, t) \binom{t}{k} (1-r)^k r^{t-k}, \end{aligned}$$

where F , k , r , and q are defined as in § 2.1. If $w > r$, then

$$(i) \quad U(f, t) = Q(f, t),$$

(ii) $s(f, t)$ is an optimal stake at time t with fortune f if and only if

$$s(f, t) \geq \max \left\{ f - F(k-1 | t-1, 1-r), \frac{r}{1-r} (F(k-1 | t-1, 1-r) - f) \right\} \\ = \underline{s}(f, t) \quad (\text{minimum bet})$$

and

$$s(f, t) \leq \min \left\{ f - F(k-2 | t-1, 1-r), \frac{r}{1-r} (F(k | t-1, 1-r) - f) \right\} \\ = \bar{s}(f, t) \quad (\text{maximum bet}).$$

Remark. As an alternate formula,

$$Q(f, t) = F(k-1 | t, 1-w) + \frac{f - F(k-1 | t, 1-r)}{\binom{t}{k} (1-r)^k r^{t-k}} \binom{t}{k} (1-w)^k w^{t-k} \\ = F(k-1 | t, 1-w) + [f - F(k-1 | t, 1-r)] \left(\frac{1-w}{1-r} \right)^k \left(\frac{w}{r} \right)^{t-k}.$$

Remark. Note that the optimal stakes do not depend on w .

To graph $U(f, t)$, just draw straight lines through the points $(F(k | t, 1-r), F(k | t, 1-w))$ when k is increasing from 0 to t . See Fig. 1.

Proof of Theorem 1. The proof follows by induction. It is trivially true for $t=0$. Assume the result for $t-1$.

First, we show that Q is obtainable ($U \geq Q$); see Cases 1 and 2, below.

Case 1. Suppose that $q(f, t) \geq k/t$, i.e., $F(k-1 | t-1, 1-r) \leq f < F(k | t, 1-r)$.

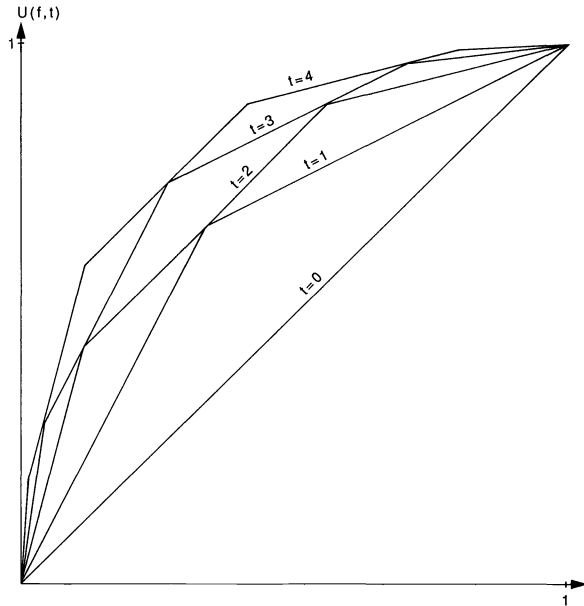


FIG. 1. Example of $U(f, t)$ with $w = \frac{2}{3}$, $r = \frac{1}{3}$, and $t = 0, 1, 2, 3$, and 4.

This case corresponds to the fortunes in Fig. 2, where the minimum stake is increasing. Let $s(f, t) = f - F(k-1|t-1, 1-r) = \underline{s}(f, t)$, the minimum stake. According to Lemma 1, f can be written as

$$f = F(k-1|t-1, 1-r) + r \frac{t}{t-k} \left(q(f, t) - \frac{k}{t} \right) \binom{t-1}{k} (1-r)^k r^{t-1-k},$$

$$\begin{aligned} U(f, t) &\geq (1-w)U(f-s, t-1) + wU\left(f + \frac{1-r}{r}s, t-1\right) \\ &= (1-w)Q(f-s, t-1) + wQ\left(f + \frac{1-r}{r}s, t-1\right) \\ &= (1-w)Q(F(k-1|t-1, 1-r), t-1) \\ &\quad + wQ\left\{F(k-1|t-1, 1-r) + \frac{t}{t-k} \left(q(f, t) - \frac{k}{t} \right) \binom{t-1}{k} (1-r)^k r^{t-1-k}, t-1\right\} \\ &= (1-w)F(k-1|t-1, 1-w) \\ &\quad + w\left\{F(k-1|t-1, 1-w) + \frac{t}{t-k} \left(q(f, t) - \frac{k}{t} \right) \binom{t-1}{k} (1-w)^k w^{t-1-k}\right\} \\ &= F(k-1|t-1, 1-w) + w \frac{t}{t-k} \left(q(f, t) - \frac{k}{t} \right) \binom{t-1}{k} (1-w)^k w^{t-1-k} \\ &= F(k-1|t, 1-w) + q(f, t) \binom{t}{k} (1-w)^k w^{t-k} = Q(f, t). \end{aligned}$$

Case 2. Suppose instead that $q(f, t) < k/t$, i.e., $f < F(k-1|t-1, 1-r)$.

Let $s(f, t) = (r/(1-r))(F(k-1|t-1, 1-r) - f) = \underline{s}(f, t)$. According to Lemma 1, f can be written as

$$f = F(k-2|t-1, 1-r) + \left(r + q(f, t) \frac{t}{k} (1-r) \right) \binom{t-1}{k-1} (1-r)^{k-1} r^{t-k},$$

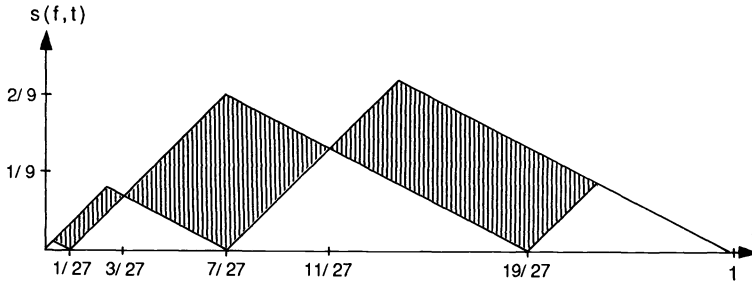


FIG. 2. Optimal stakes $s(f, t)$ when $r = \frac{1}{3}$ and $t = 4$.

$$\begin{aligned}
U(f, t) &\geq (1-w)Q(f-s, t-1) + wQ\left(f + \frac{1-r}{r}s, t-1\right) \\
&= (1-w)Q\left\{F(k-2|t-1, 1-r) + \frac{t}{k}q(f, t)\binom{t-1}{k-1}(1-r)^{k-1}r^{t-k}, t-1\right\} \\
&\quad + wQ(F(k-1|t-1, 1-r), t-1) \\
&= (1-w)\left\{F(k-2|t-1, 1-w) + q(f, t)\binom{t}{k}(1-w)^{k-1}w^{t-k}\right\} \\
&\quad + wF(k-1|t-1, 1-w) \\
&= F(k-1|t, 1-w) + q(f, t)\binom{t}{k}(1-w)^kw^{t-k} = Q(f, t).
\end{aligned}$$

Hence Q is obtainable.

Now we show that we cannot do any better than Q ($U \leq Q$). See Fig. 1 throughout the rest of the proof.

Let

$$\begin{aligned}
E_s(f, t) &= (1-w)U(f-s, t-1) + wU\left(f + \frac{1-r}{r}s, t-1\right) \\
&= \text{constant} + w\frac{1-r}{r}s\left(\frac{1-w}{1-r}\right)^n\left(\frac{w}{r}\right)^{t-1-n} - (1-w)s\left(\frac{1-w}{1-r}\right)^m\left(\frac{w}{r}\right)^{t-1-m},
\end{aligned}$$

where m is such that $F(m-1|t-1, 1-r) \leq f-s < F(m|t-1, 1-r)$, and n is such that $F(n-1|t-1, 1-r) \leq f + ((1-r)/r)s < F(n|t-1, 1-r)$. m and n depend on s .

Note that $E_s(f, t)$ is continuous and almost everywhere differentiable with respect to s . We know that $U(f, t) = \max_s E_s(f, t)$ [2].

To prove that s is an optimal bet if and only if $\underline{s} \leq s \leq \bar{s}$ and that we cannot do any better than Q , we show that

$$\frac{d}{ds} E_s(f, t) \begin{cases} > 0 & \text{when } s < \underline{s}, \\ = 0 & \text{when } \underline{s} < s < \bar{s}, \\ < 0 & \text{when } s > \bar{s} \end{cases}$$

for all s , where E_s is differentiable.

If $s < \underline{s}$, then $m = n$; so

$$\frac{d}{ds} E_s(f, t) = w\left(\frac{1-w}{1-r}\right)^n\left(\frac{w}{r}\right)^{t-1-n}\left(\frac{1-r}{r} - \frac{1-w}{w}\right) > 0 \quad \text{since } w > r.$$

If $\underline{s} < s < \bar{s}$, then $m = n-1$;

$$\text{so } \frac{d}{ds} E_s(f, t) = \frac{(1-w)^nw^{t-n}}{(1-r)^{n-1}r^{t-n}} - \frac{(1-w)^nw^{t-n}}{(1-r)^{n-1}r^{t-n}} = 0.$$

If $s > \bar{s}$, then $m \leq n-2 = n-1-i$ for some positive integer i ; so

$$\begin{aligned}
\frac{d}{ds} E_s(f, t) &= \frac{(1-w)^{n-1}w^{t-n}}{(1-r)^{n-1}r^{t-n}} - \frac{(1-w)^{n-i}w^{t-n+i}}{(1-r)^{n-1-i}r^{t-n+i}} \\
&= \frac{(1-w)^{n-i}w^{t-n}}{(1-r)^{n-1-i}r^{t-n}} \left(\left(\frac{1-w}{1-r}\right)^i - \left(\frac{w}{r}\right)^i \right) < 0 \quad \text{since } w > r \text{ and } i \geq 1. \quad \square
\end{aligned}$$

2.4. A 0/1 utility function. In this section we consider the case where $u(f) = 0$ for $f < 1$ and $u(f) = 1$ for $f = 1$. $U(f, t)$ is, as previously defined, the probability of reaching 1 using an optimal strategy and starting at fortune f with t times left to play.

DEFINITION. A fortune f is binomial at time t if $q(f, t) \binom{t}{k}$ is an integer.

Examples. (i) The binomial fortunes when $t=2$ are 0, r^2 , $r^2 + r(1-r) = r$, $r^2 + 2r(1-r) = 1 - (1-r)^2$, and 1.

(ii) When $r = \frac{1}{2}$, then the binomial fortunes at time t are the integer multiples of $\frac{1}{2^t}$, i.e., the previously defined binary fortunes.

DEFINITION. A stake $s(f, t)$ is binomial if we arrive at a binomial fortune, regardless of whether we win or lose, i.e., if $f - s(f, t)$ and $f + ((1-r)/r)s(f, t)$ are binomial fortunes at time $t-1$.

THEOREM 2. If $w > r$ and f is binomial, then $U(f, t) = Q(f, t)$ (as in Theorem 1), and $s(f, t)$ is optimal if and only if

- (i) $s(f, t) \geq \max \{f - F(k-1 | t-1, 1-r), (r/(1-r))(F(k-1 | t-1, 1-r) - f)\} \triangleq \underline{s}(f, t),$
- (ii) $s(f, t) \leq \min \{f - F(k-2 | t-1, 1-r), (r/(1-r))(F(k | t-1, 1-r) - f)\} \triangleq \bar{s}(f, t),$
- (iii) $s(f, t)$ is a binomial stake.

Compare this with Proposition 4 and Theorem 1. It means that for binomial f we can do just as well for this case as for the case of § 2.3, despite $u(f)$ being smaller.

Proof. The proof follows by induction. It is trivially true for $t=0$. We only must show that there always exists a binomial stake fulfilling conditions (i) and (ii). The rest follows from Theorem 1.

Case 1. Suppose that $q(f, t) \geq k/t$. We want to show that $s(f, t) = f - F(k-1 | t-1, 1-r)$ ($= \underline{s}(f, t)$) is a binomial stake.

- (i) $f - s = F(k-1 | t-1, 1-r)$, which is binomial at time $t-1$,
- (ii)

$$\begin{aligned} f + \frac{1-r}{r}s &= F(k-1 | t-1, 1-r) + \frac{t}{t-k} \left(q(f, t) - \frac{k}{t} \right) \binom{t-1}{k} (1-r)^k r^{t-1-k} \\ &= F(k-1 | t-1, 1-r) + q(f, t) \binom{t}{k} (1-r)^k r^{t-1-k}, \end{aligned}$$

which is binomial at time $t-1$, since f being binomial at time t implies that $q(f, t) \binom{t}{k}$ is an integer.

Case 2. Suppose instead that $q(f, t) < k/t$. Now we want to show that $s(f, t) = (r/(1-r))(F(k-1 | t-1, 1-r) - f)$ ($= \bar{s}(f, t)$) is a binomial stake.

- (i) $f + ((1-r)/r)s = F(k-1 | t-1, 1-r)$, which is binomial at time $t-1$,
- (ii)

$$\begin{aligned} f - s + F(k-2 | t-1, 1-r) + \frac{t}{k} q(f, t) \binom{t-1}{k-1} (1-r)^{k-1} r^{t-k} \\ = F(k-2 | t-1, 1-r) + q(f, t) \binom{t}{k} (1-r)^{k-1} r^{t-k}, \end{aligned}$$

which is binomial at time $t-1$, since $q(f, t) \binom{t}{k}$ is integer-valued. \square

For nonbinomial f , we can use the fact that U is nondecreasing to obtain a lower bound for U , and we can use Theorem 1 to obtain an upper bound. Let $\tilde{Q}(f, t) = F(k-1 | t, 1-w) + [q(f, t) \binom{t}{k}] (1-w)^k w^{t-k}$, where $f = F(k-1 | t, 1-r) + q(f, t) \binom{t}{k} \cdot (1-r)^k r^{t-k}$ ($[x]$ is the integer part of x). This means that \tilde{Q} takes jumps of magnitudes $(1-w)^k w^{t-k}$ (k varies) at binomial fortunes, and is constant otherwise. For binomial fortunes, $\tilde{Q} = U$.

COROLLARY 2. It holds that $\tilde{Q}(f, t) \leq U(f, t) \leq Q(f, t)$.

Since $|Q(f, t) - \tilde{Q}(f, t)|$ converges to zero when $t \rightarrow \infty$, the bounds will be good for large t . For the case of $r \leq \frac{1}{2} \leq w$, the result is stronger.

THEOREM 3. *If $r \leq \frac{1}{2} \leq w$, then $U(f, t) = \tilde{Q}(f, t)$.*

It is easy to find a counterexample to this equality for general r and w . Just pick any r and w such that $r < w < \frac{1}{2}$ or $\frac{1}{2} < r < w$ and try it for sufficiently large t .

Proof. The proof follows by induction. It is trivially true for $t = 0$. Assume the result for $t - 1$. Let f_1, f_2, \dots and g_1, g_2, \dots be the binomial fortunes at time t and $t - 1$, respectively. We know that $g_i - g_{i-1} = (1 - r)^k r^{t-1-k}$ for some k and that $U(g_i, t - 1) - U(g_{i-1}, t - 1) = (1 - w)^k w^{t-1-k}$ for the same k .

Since U is nondecreasing, and since $\tilde{Q}(f, t)$ is a step function, it is enough to show that $U(f, t) = \tilde{Q}(f, t)$ at fortunes $f_j - \varepsilon$, where $\varepsilon > 0$ is arbitrarily small. To be more precise, we must show that $U(f_j - \varepsilon, t) = U(f_{j-1}, t)$ for all binomial fortunes f_j . It is clear that an optimal stake can be found among those of the form $f_j - \varepsilon - g_i \geq 0$. So

$$U(f_j - \varepsilon, t) = \max_i \left\{ (1 - w)U(g_i + \varepsilon, t - 1) + wU\left(f_j + \frac{1 - r}{r}(f_j - \varepsilon - g_i), t - 1\right) \right\} \triangleq \max_i U_i.$$

Let $g_x = f_j - \bar{s}$, $g_m = f_j - \underline{s}$, $g_{\bar{x}} = f_j + ((1 - r)/r)\bar{s}$, and $g_{\bar{m}} = f_j + ((1 - r)/r)\underline{s}$, where x and m are index variables.

Let us first try $g_x \leq g_i \leq g_m$, below:

$$U_i = (1 - w)U(g_i + \varepsilon, t - 1) + wU(g_{\bar{x}-(i-x)} - o(\varepsilon), t - 1),$$

so

$$\begin{aligned} U_i &= (1 - w)U(g_i, t - 1) + wU(g_{\bar{x}-(i-x)-1}, t - 1) \\ &= (1 - w)U(g_i, t - 1) + wU(g_{\bar{x}-(i-x)}, t - 1) - w(1 - w)^k w^{t-1-k} \\ &= U(f_j, t) - (1 - w)^k w^{t-k} = U(f_{j-1}, t). \end{aligned}$$

Now let us try $g_i < g_x$, below:

$$U_i = (1 - w)U(g_i + \varepsilon, t - 1) + wU(g, t - 1),$$

where $g < g_{\bar{x}+(x-i)}$ since $r \leq \frac{1}{2}$; so

$$\begin{aligned} U_i &\leq (1 - w)U(g_i, t - 1) + wU(g_{\bar{m}-(i-m)-1}, t - 1) \\ &\leq (1 - w)\{U(g_x, t - 1) - (x - i)(1 - w)^{k-1} w^{t-1-(k-1)}\} \\ &\quad + w\{U(g_{\bar{x}}, t - 1) + (x - i - 1)(1 - w)^{k+1} w^{t-1-(k+1)}\} \\ &= U(f_j, t) - (x - i - 1) \\ &\quad \cdot \{(1 - w)^k w^{t-k} - (1 - w)^{k+1} w^{t-(k+1)}\} - (1 - w)^k w^{t-k} \\ &\leq U(f_j, t) - (1 - w)^k w^{t-k} \text{ (since } w \geq \frac{1}{2}) = U(f_{j-1}, t). \end{aligned}$$

It remains to try $g_i > g_m$, below:

$$U_i = (1 - w)U(g_i, t - 1) + wU(g, t - 1),$$

where $g < g_{\bar{m}-(i-m)}$ since $r \leq \frac{1}{2}$; so

$$\begin{aligned} U_i &\leq (1 - w)U(g_i, t - 1) + wU(g_{\bar{m}-(i-m)-1}, t - 1) \\ &= (1 - w)\{U(g_m, t - 1) + (i - m)(1 - w)^k w^{t-1-k}\} \\ &\quad + w\{U(g_{\bar{m}}, t - 1) - (i - m - 1)(1 - w)^k w^{t-1-k}\} \\ &= U(f_j, t) - (i - m)\{(1 - w)^k w^{t-1-k}(w - (1 - w))\} - (1 - w)^k w^{t-k} \\ &\leq U(f_j, t) - (1 - w)^k w^{t-k} \text{ (since } w \geq \frac{1}{2}) = U(f_{j-1}, t). \end{aligned}$$

□

To know which bets are optimal ($r \leq \frac{1}{2} \leq w$), we divide our fortune f into two parts:

- (i) $f - \{q(f, t) \binom{t}{k} - [q(f, t) \binom{t}{k}]\}(1-w)^k w^{t-k}$, the valuable part;
- (ii) $\{q(f, t) \binom{t}{k} - [q(f, t) \binom{t}{k}]\}(1-w)^k w^{t-k}$, the worthless part.

If we use the valuable part and the money it generates during the course of the play, to bet, according to Theorem 2, we achieve optimum. This means that we can use the worthless part and the money it generates for bets whenever we want, and the strategy will not influence our chances to reach our goal. We could, of course, also use it, before we start playing, for some luck-bringing endeavor, such as throwing it in a fountain. The strategies described here are not the only optimal ones, except when the worthless part is zero (i.e., f is binomial).

2.5. Other utility functions. The results of this section follow immediately from the previous theorems. Let $S_1(f, t)$ and $S_3(f, t)$ be the sets of optimal stakes in Theorems 1 and 3, respectively.

COROLLARY 3. *If $u(f) = Q(f, s)$ (or $\tilde{Q}(f, s)$), then $U(f, t) = Q(f, s+t)$ (respectively, $\tilde{Q}(f, s+t)$), and $s(f, t)$ is optimal if it belongs to the set $S_1(f, s+t)(S_3(f, s+t))$.*

2.6. Limit results and approximation formulas. All limits in this section are the same for the cases of §§ 2.3 and 2.4, so we do not make a distinction. Φ denotes the cumulative distribution function of the standard normal distribution.

What happens to $U(f, t)$ when $t \rightarrow \infty$? It is easy to see that it converges to 1 pointwise. However, this is not the limit that we are interested in. We want to know the limit of $U(f, t)$ when w and r vary with t in such a way that $w - r \rightarrow 0$ as $t \rightarrow \infty$. This means that the bets get less favorable, but this is compensated by the number of bets allowed, which increase. We do this in such a way that $\sqrt{t}(w - r)$ equals a constant c for all t .

THEOREM 4. *If $\sqrt{t}(w - r) = c$ and $w \rightarrow w_0 \in (0, 1)$, then $\lim_{t \rightarrow \infty} U(f, t) = \Phi(\Phi^{-1}(f) + c/\sqrt{w_0(1-w_0)})$.*

We obtain this same function as U for similar continuous time problems in Theorems 6 and 7.

Proof. It holds that

$$\begin{aligned} f &< F(k(f, t) | t, 1-r) = P(S_t \leq k(f, t)) \\ &= P\left(\frac{S_t - t(1-r)}{\sqrt{tr(1-r)}} \leq \frac{k(f, t) - t(1-r)}{\sqrt{tr(1-r)}}\right) = \Phi\left(\frac{k(f, t) - t(1-r)}{\sqrt{tr(1-r)}}\right) + \Delta_t, \end{aligned}$$

where $S_t \sim \text{Bin}(t, 1-r)$ and where $\lim_{t \rightarrow \infty} \Delta_t = 0$, since $(k(f, t) - t(1-r))/\sqrt{tr(1-r)}$ converges. Hence $k(f, t) > \Phi^{-1}(f - \Delta_t)\sqrt{tr(1-r)} + t(1-r)$. Likewise,

$$f \geq F(k(f, t) - 1 | t, 1-r) = \Phi\left(\frac{k(f, t) - 1 - t(1-r)}{\sqrt{tr(1-r)}}\right) + \Delta'_t;$$

so $k(f, t) \leq \Phi^{-1}(f - \Delta'_t)\sqrt{tr(1-r)} + t(1-r) + 1$. Now

$$\begin{aligned} U(f, t) &< F(k(f, t) | t, 1-w) = P(R_t \leq k(f, t)) \\ &= P\left(\frac{R_t - t(1-w)}{\sqrt{tw(1-w)}} \leq \frac{k(f, t) - t(1-w)}{\sqrt{tw(1-w)}}\right) = \Phi\left(\frac{k(f, t) - t(1-w)}{\sqrt{tw(1-w)}}\right) + \varepsilon_t \\ &\leq \Phi\left(\frac{\Phi^{-1}(f - \Delta'_t)\sqrt{tr(1-r)} + t(1-r) + 1 - t(1-w)}{\sqrt{tw(1-w)}}\right) + \varepsilon_t \\ &= \Phi\left(\Phi^{-1}(f - \Delta'_t)\sqrt{\frac{r(1-r)}{w(1-w)}} + \frac{c}{\sqrt{w(1-w)}} + \frac{1}{\sqrt{tw(1-w)}}\right) + \varepsilon_t, \end{aligned}$$

where $R_t \sim \text{Bin}(t, 1-w)$ and where $\lim_{t \rightarrow \infty} \varepsilon_t = 0$. Likewise,

$$\begin{aligned} U(f, t) &\geq F(k(f, t) - 1 | t, 1-w) + \Phi\left(\frac{k(f, t) - 1 - t(1-w)}{\sqrt{tw(1-w)}}\right) + \varepsilon'_t \\ &> \Phi\left(\Phi^{-1}(f - \Delta_t) \sqrt{\frac{r(1-r)}{w(1-w)}} + \frac{c}{\sqrt{w(1-w)}} - \frac{1}{\sqrt{tw(1-w)}}\right) + \varepsilon'_t; \end{aligned}$$

so $\lim_{t \rightarrow \infty, w \rightarrow w_0} U(f, t) = \Phi(\Phi^{-1}(f) + c/\sqrt{w_0(1-w_0)})$. \square

If we want to use this result to provide approximate values for $U(f, t)$, we should perhaps use the following corollary instead, which follows from the proof of the theorem.

COROLLARY 4. *If t , $tw(1-w)$ and $tr(1-r)$ are all large, then*

$$U(f, t) \approx \Phi\left(\Phi^{-1}(f) \sqrt{\frac{r(1-r)}{w(1-w)}} + \frac{\sqrt{t(w-r)}}{\sqrt{w(1-w)}}\right).$$

We might also be interested in the limit of $s(f, t)$ as $t \rightarrow \infty$. It is not difficult to see that it converges to zero uniformly (also when $w-r \rightarrow 0$), but this does not relay anything about how fast or in what way it converges. Neither does it give us an approximation formula for the bet sizes. Betting zero all the time is gainless.

Instead, we consider the limit of $\sqrt{t} s(f, t)$ as $t \rightarrow \infty$. Since $s(f, t)$ is not unique, we can choose different sequences of $\sqrt{t} s(f, t)$, some of which may not converge. Which sequence should we choose? When $f = F(k | t, 1-r)$, the bet size is unique, so then the choice is obvious. For $F(k-1 | t, 1-r) < f < F(k | t, 1-r)$, we choose a bet such that $s(f, t)$ is between $s(F(k-1 | t, 1-r), t)$ and $s(F(k | t, 1-r), t)$. In the case of § 2.3, there is always such an optimal bet. That is also true for the case of § 2.4, with the exception of when $k = t/2$.

THEOREM 5. *It holds that $\lim_{t \rightarrow \infty} \sqrt{t} s(f, t) = \sqrt{(r/(1-r))} \phi(\Phi^{-1}(f))$, where ϕ is the density function of the standard normal distribution.*

Proof. It is enough to consider fortunes of the form $F(k-1 | t, 1-r)$. Let

$$\begin{aligned} f &= F(k-1 | t, 1-r) \text{ so } s(f, t) = F(k-1 | t, 1-r) - F(k-2 | t-1, 1-r) \\ &= (1-r)F(k-2 | t-1, 1-r) + rF(k-1 | t-1, 1-r) - F(k-2 | t-1, 1-r) \\ &= F(k-1 | t-1, 1-r) - F(k-2 | t-1, 1-r); \end{aligned}$$

$k(f, t) > \Phi^{-1}(f - \Delta_t) \sqrt{tr(1-r)} + t(1-r)$ and $k(f, t) \leq \Phi^{-1}(f - \Delta'_t) \sqrt{tr(1-r)} + t(1-r) + 1$; so $k(f, t) = \Phi^{-1}(f - \tilde{\Delta}_t) \sqrt{tr(1-r)} + t(1-r) + d$, where $d \in [0, 1]$. It holds that

$$\begin{aligned} \lim_{t \rightarrow \infty} \sqrt{t} s(f, t) &= \lim_{t \rightarrow \infty} \sqrt{t} r \{F(k-1 | t-1, 1-r) - F(k-2 | t-1, 1-r)\} \\ &= r \lim_{t \rightarrow \infty} \sqrt{t} P(S_{t-1} = k-1) \\ &= r \lim_{t \rightarrow \infty} \sqrt{t} \left(\frac{1}{\sqrt{tr(1-r)}} \phi\left(\frac{k(f, t) - 1 - t(1-r)}{\sqrt{tr(1-r)}}\right) + \varepsilon_t \right) \\ &= \sqrt{\frac{r}{1-r}} \lim_{t \rightarrow \infty} \phi\left(\Phi^{-1}(f + \tilde{\Delta}_t) + \frac{d-1}{\sqrt{tr(1-r)}}\right) + r \lim_{t \rightarrow \infty} \sqrt{t} \varepsilon_t \\ &= \sqrt{\frac{r}{1-r}} \phi(\Phi^{-1}(f)), \end{aligned}$$

since

$$\begin{aligned} \left| r \lim_{t \rightarrow \infty} \sqrt{t} \varepsilon_t \right| &\leq r \lim_{t \rightarrow \infty} \sqrt{t} \left| \frac{1}{\sqrt{tr(1-r)}} \left(\frac{A}{t} + \frac{B}{\sqrt{t}} \left(\frac{k-t(1-r)}{\sqrt{t} r(1-r)} \right)^2 \right) \right| \\ &= \sqrt{\frac{r}{1-r}} \lim_{t \rightarrow \infty} \left| \frac{A}{t} + \frac{B}{\sqrt{t}} \right| \left| \left(\Phi^{-1}(f - \tilde{\Delta}_t) + \frac{d}{\sqrt{tr(1-r)}} \right)^2 \right| = 0, \end{aligned}$$

where A and B are constants. For the bound on ε_t , see Feller [3, p. 170]. \square

COROLLARY 5. *If t and $tr(1-r)$ are large, then $s(f, t) \approx (1/\sqrt{t}) \cdot \sqrt{r/(1-r)} \phi(\Phi^{-1}(f))$.*

For $r = \frac{1}{2}$, this is similar to the optimal stake for the continuous time problem with constant drift μ_t (Theorems 6 and 7).

The variance of a single bet is $s(f, t)^2(w(1-w)/r^2)$. If we multiply the stake by $\sqrt{r/(1-r)}$, we get variance $(s(f, t) \sqrt{r/(1-r)})^2(w(1-w)/r^2) = s(f, t)^2(w(1-w)/r(1-r))$, which is equal to $s(f, t)^2$ in the limit, since $w-r \rightarrow 0$. This is the intuition behind the $\sqrt{r/(1-r)}$ part of the above formulas.

3. Continuous time problem. Consider the stochastic process X_t such that $dX_t = \sigma_t(\mu_t dt + dB_t)$, where B_t is the standard Brownian motion. In the gambling context, X_t represents the fortune at time t . The fortune space is of the form $[a, b]$, and without loss of generality, we assume it to be $[0, 1]$. μ_t is a fixed function and, for some constant M , $\mu_t \leq M$ for all t . Zero and 1 are absorbing states, so, if $X_t = 0$ or 1, then $\sigma_s = 0$ for all $s \geq t$. Otherwise, the control variable σ_t , which may only depend on what has happened up to time t , can be chosen to be any nonnegative value as long as

$$\int_t^{T-\varepsilon} |\sigma_s|^2 ds < \infty \quad \forall \varepsilon > 0 \quad \left(\text{which implies } \int_t^{T-\varepsilon} |\mu_s \sigma_s| ds < \infty \right).$$

T is a fixed stopping time, and our goal is to maximize the expected value of a utility function $u(f, T)$ which expresses the value of having fortune f at time T . In addition, we want to know $U(f, t)$; the expected utility at time t , given that we have fortune f at time t and that we use an optimal strategy.

With Φ we denote the cumulative distribution function of the standard normal distribution, and with ϕ its density function. Let $\sigma^*(f, t)$ be the optimal choice of σ_t if we have fortune f at time t .

Let $S = \{(f, t) : 0 \leq f \leq 1, 0 \leq t \leq T\}$ be the state space. Let $\mu_t^+ = \max\{0, \mu_t\}$.

DEFINITION. Let $m_t = \sqrt{\int_t^T (\mu_s^+)^2 ds}$, a measure of the remaining amount of “favorability.” Note that when $\mu_t = \mu$, for all t , then $m_t = \sqrt{T-t} \mu$.

Without loss of generality, we assume that $m_t > 0$ when $t < T$. This property can easily be obtained by a change of T , since $\sigma_t = 0$ is one of the optimal strategies when $m_t = 0$.

3.1. Utility function $u(f, T) = f$. We first study the problem with the utility function $u(f, T) = f$.

THEOREM 6. *If $u(f, T) = f$, then $U(f, t) = \tilde{U}(f, t) \triangleq \Phi(\Phi^{-1}(f) + m_t)$, $(f, t) \in S$ and $\sigma^*(f, t) = \tilde{\sigma}^*(f, t) \triangleq \phi(\Phi^{-1}(f))(\mu_t^+/m_t)$, $0 \leq f \leq 1$, $0 \leq t < T$.*

Remark. For $\mu_t = \mu$, for all t , and for $r = T-t$, the time left to play, we get the formulas

$$U(f, T-r) = \Phi(\Phi^{-1}(f) + \sqrt{r} \mu) \quad \text{and} \quad \sigma^*(f, T-r) = \frac{1}{\sqrt{r}} \phi(\Phi^{-1}(f)).$$

See Figs. 3 and 4.

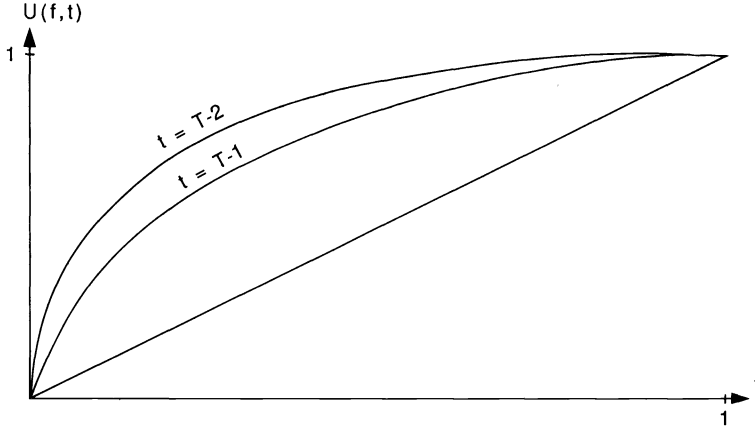


FIG. 3. $U(f, t)$ when μ_t is a constant equal to 1 and there are one and two time units left to play ($m_t = 1$ and $\sqrt{2}$).

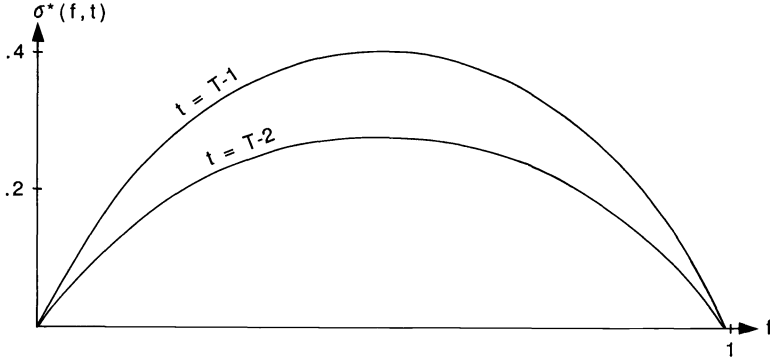


FIG. 4. $\sigma^*(f, t)$ when μ_t is a constant and there are one and two time units left to play ($\mu_t^+ / m_t = 1$ and $1/\sqrt{2}$).

Note that this result is the same as that for the limit of the discrete time problem of § 2.6. Note also that the bet size does not depend on μ in this special case. In the general case, the bet size only depends on μ_τ in relation to m_t . That is, for two functions μ_t and $\bar{\mu}_t$, the optimal control is the same if $\bar{\mu}_t = C\mu_t$, for all t and some positive constant C .

The main tool in proving the theorem is Ito's formula. Since the derivatives of U are not finite, we are not able to use the formula directly. Instead, we first must look at a problem where the state space is reduced to $S_\varepsilon = \{(f, t): \varepsilon \leq f \leq 1, 0 \leq t \leq T - \varepsilon\}$ and where $u(\varepsilon, t) = \tilde{U}(\varepsilon, t)$, $u(f, T - \varepsilon) = \tilde{U}(f, T - \varepsilon)$ and $u(f, t) = 0$, otherwise. $u(f, \tau)$ is the utility function at fortune f if we stop the process at time τ . Let U_ε and σ_ε^* be the U and σ^* functions of this modified problem.

LEMMA 2. It holds that

$$U_\varepsilon(f, t) = \tilde{U}(f, t) \quad \forall (f, t) \in S_\varepsilon \quad \text{and} \quad \sigma_\varepsilon^*(f, t) = \tilde{\sigma}^*(f, t) \quad \forall (f, t) \in \text{interior of } S_\varepsilon.$$

Proof. Let $U_\varepsilon^t(U_\varepsilon^f)$ be the derivative of U_ε with respect to time (fortune); let U_ε^{ff} be the second derivative of U_ε with respect to fortune; and let $I(\sigma) = U_\varepsilon^t(X_s, s) + U_\varepsilon^f(X_s, s)\sigma_s\mu_s + \frac{1}{2}U_\varepsilon^{ff}(X_s, s)\sigma_s^2$. By definition, $U_\varepsilon(f, t) = \tilde{U}(f, t)$ on the boundaries of S_ε . First, we must show that $\tilde{U}(X_t, t)$ is a martingale for the strategy σ_ε^* (\tilde{U} is obtainable). Then we must show that $\tilde{U}(X_t, t)$ is a supermartingale for all

strategies (\tilde{U} is excessive). Let τ be the stopping time when X_t hits the boundary of S_ε . We note that $U \in C^2$ on an open set containing S_ε and that

$$E\left(\int_t^{T-\varepsilon} |\mu_s \sigma_s^*| ds\right) \leq \int_t^{T-\varepsilon} M\sigma^*\left(\frac{1}{2}, T-\varepsilon\right) ds < \infty$$

and

$$E\left(\int_t^{T-\varepsilon} |\sigma_s^*|^2 ds\right) \leq \int_t^{T-\varepsilon} \left|\sigma^*\left(\frac{1}{2}, T-\varepsilon\right)\right|^2 ds < \infty.$$

Hence we can use Ito's formula as follows:

$$\begin{aligned} E[U_\varepsilon(x_\tau, \tau) | X(t)] &= E\left[U_\varepsilon(X_t, t) + \int_t^\tau I(\sigma) ds + \int_t^\tau U_\varepsilon^f(x_s, s) \sigma_s dB_s \middle| X_t\right] \\ &= U_\varepsilon(X_t, t) + E\left[\int_t^\tau I(\sigma) ds \middle| X_t\right] \\ &\begin{cases} = U_\varepsilon(X_t, t) \text{ if } I(\sigma) = 0, \text{ which makes } U_\varepsilon(X_t, t) \text{ a martingale,} \\ \leq U_\varepsilon(X_t, t) \text{ if } I(\sigma) \leq 0, \text{ which makes } U_\varepsilon(X_t, t) \text{ a supermartingale,} \end{cases} \end{aligned}$$

since the expected value of the integral with respect to B_s is zero. So it only remains to show that $I(\sigma^*) = 0$ and that $I(\sigma) \leq 0$ for all σ ; see the following equations:

$$\begin{aligned} I(\sigma^*) &= -\phi(\Phi^{-1}(f) + m_s) \frac{(\mu_s^+)^2}{2m_s} + \frac{\phi(\Phi^{-1}(f) + m_s)}{\phi(\Phi^{-1}(f))} \phi(\Phi^{-1}(f)) \frac{\mu_s^+}{m_s} \mu_s^+ \\ &\quad + \frac{1}{2} (-1) \frac{\phi(\Phi^{-1}(f)) m_s}{\phi(\Phi^{-1}(f))^2} \phi(\Phi^{-1}(f))^2 \frac{(\mu_s^+)^2}{m_s^2} \\ &= \phi(\Phi^{-1}(f) + m_s) \frac{(\mu_s^+)^2}{m_s} \left(-\frac{1}{2} + 1 - \frac{1}{2}\right) = 0; \\ 0 &= \frac{dI(\sigma)}{d\sigma} = U_\varepsilon^f(X_s, s) \mu_s^+ + U_\varepsilon^g(X_s, s) \sigma_s \\ &= \frac{\phi(\Phi^{-1}(f) + m_s)}{\phi(\Phi^{-1}(f))} \mu_s^+ + (-1) \frac{\phi(\Phi^{-1}(f) + m_s) m_s}{\phi(\Phi^{-1}(f))^2} \sigma_s \\ &= \frac{\phi(\Phi^{-1}(f) + m_s)}{\phi(\Phi^{-1}(f))} \left(\mu_s^+ - \frac{m_s \sigma_s}{\phi(\Phi^{-1}(f))^2}\right), \end{aligned}$$

which implies that $\sigma_s = \phi(\Phi^{-1}(f))(\mu_s^+/m_s) = \sigma_s^*$, which maximizes $I(\sigma)$, since

$$\frac{d^2 I(\sigma)}{d\sigma^2} = U_\varepsilon^g(X_s, s) = -\frac{\phi(\Phi^{-1}(f) + m_s)}{\phi(\Phi^{-1}(f))^2} m_s < 0. \quad \square$$

Proof of Theorem 6. Since $\tilde{U}(f, T) = U(f, T) = u(f, T)$ and $\tilde{U}(0, t) = U(0, t) = u(0, t)$, and since \tilde{U} and U are bounded continuous functions, there exists a $\delta_\varepsilon > 0$ for every $\varepsilon > 0$ such that

- (i) $\lim_{\varepsilon \rightarrow 0} \delta_\varepsilon = 0$;
- (ii) $|U(f, t) - u(0, t)| < \delta_\varepsilon$ and $|\tilde{U}(f, t) - u(0, t)| < \delta_\varepsilon$, for all $f \leq \varepsilon$;
- (iii) $|U(f, t) - u(f, T)| < \delta_\varepsilon$ and $|\tilde{U}(f, t) - u(f, T)| < \delta_\varepsilon$, for all $t \geq T - \varepsilon$.

Take any $(f, t) \in S$. We want to show that $U(f, t) = \tilde{U}(f, t)$. On the boundaries of S , it is true by definition. On the interior, we have that, for sufficiently small ε (Lemma 2),

$$\begin{aligned} U(f, t) - \tilde{U}(f, t) &= U(f, t) - U_\varepsilon(f, t) \\ &= \lim_{\varepsilon \rightarrow 0} (U(f, t) - U_\varepsilon(f, t)) \\ &\geq \lim_{\varepsilon \rightarrow 0} \left(E \int U(f_\tau, \tau) dv - E \int U_\varepsilon(f_\tau, \tau) dv \right) \\ &= \lim_{\varepsilon \rightarrow 0} E \left[\int U(f_\tau, \tau) - u_\varepsilon(f_\tau, \tau) dv \right] \geq \lim_{\varepsilon \rightarrow 0} E \left[\int -2\delta_\varepsilon dv \right] = 0, \end{aligned}$$

where τ is as before and v is the probability measure of (f_τ, t) under the strategy σ_ε^* . Likewise,

$$\begin{aligned} U(f, t) - \tilde{U}(f, t) &\leq \lim_{\varepsilon \rightarrow 0} E \int U(f_\tau, \tau) dw - E \int U_\varepsilon(f_\tau, \tau) dw \\ &= \lim_{\varepsilon \rightarrow 0} E \left[\int U(f_\tau, \tau) - U_\varepsilon(f_\tau, \tau) dw \right] \leq \lim_{\varepsilon \rightarrow 0} E \int 2\delta_\varepsilon dw = 0, \end{aligned}$$

where w is the probability measure of (f_τ, τ) under the strategy σ^* . Hence $U(f, t) = \tilde{U}(f, t)$ on the interior of S , which implies that $\sigma^* = \sigma_\varepsilon^*$ on the interior of S . \square

3.2. The 0/1 utility function. We now shift our attention to the utility function $u(f, T) = 0$ for $f < 1$ and $u(1, T) = 1$. This is the one for which Sudderth and Weerasinghe [6] solved the subfair case.

THEOREM 7. *If $u(f, T) = 0$ for $f < 1$ and $u(1, T) = 1$, then*

$$U(f, t) = \tilde{U}(f, t) \triangleq \Phi(\Phi^{-1}(f) + m_t), \quad (f, t) \in S$$

and

$$\sigma^*(f, t) = \tilde{\sigma}^*(f, t) \triangleq \phi(\Phi^{-1}(f))(\mu_t^+ / m_t), \quad 0 \leq f \leq 1, 0 \leq t < T.$$

Proof. Since, for all f , the utility function $u(f, T)$ is smaller than that in § 3.1, we know from Theorem 6 that $U(f, t) \leq \tilde{U}(f, t)$.

To show that $U(f, t) \geq \tilde{U}(f, t)$, we consider the following modified problem. Let $u'_t = \mu_{t+\varepsilon}$ for $t \leq T - \varepsilon$ and $u'_t = 0$ for $T - \varepsilon < t \leq T$. Consider the strategy $\sigma(f, t) = 1/(T - t)$ for $T \geq T - \varepsilon$. With it, we leave the interval $(0, 1)$ with probability 1 before time T , since $\int_{T-\varepsilon}^T |1/(T-s)|^2 ds = \infty$. The probability of leaving the interval at $f = 1$ is equal to the fortune at time $T - \varepsilon$, since there is no drift ($u'_t = 0$). Hence $U_\varepsilon(f, t) \geq f$ for $t \geq T - \varepsilon$, where U_ε is the “ U -function” of the modified problem.

From this and Theorem 6, we have that $U_\varepsilon(f, t) \geq \tilde{U}(f, t + \varepsilon)$. Now $U(f, t) \geq U_\varepsilon(f, t)$ for all ε , and hence $U(f, t) \geq \lim_{\varepsilon \rightarrow 0} U_\varepsilon(f, t) \geq \lim_{\varepsilon \rightarrow 0} \tilde{U}(f, t + \varepsilon) = \tilde{U}(f, t)$. $\sigma^*(f, t) = \tilde{\sigma}^*(f, t)$ now follows from Theorem 6. \square

Remark. In this paper we have fixed μ_t as a predetermined function, while the control variable $\sigma(f, t)$ can be chosen almost freely. If we also have μ_t as a control variable, i.e., we can choose from among various μ_t -functions, then it follows from Theorems 6 and 7, respectively, that we should pick the function that maximizes m_t , while $\sigma(f, t)$ is chosen as before. It is an open question regarding what happens if there is a restriction on the choice of $\sigma(f, t)$, such as an upper limit, or if the allowable controls depend on the present fortune.

3.3. Other utility functions.

COROLLARY 6 (to Theorem 6). *If $u(f, T) = \Phi(\Phi^{-1}(f) + m)$, $m \geq 0$, then*

$$U(f, t) = \Phi(\Phi^{-1}(f) + m + m_t) \quad \text{and} \quad \sigma^*(f, t) = \phi(\Phi^{-1}(f))\mu_t^+ / (m + m_t).$$

Remark. Note that the optimal strategy is no longer independent of μ when $\mu_t = \mu$ for all t . We assume that the intuition behind this is that $u(f, T)$ is no longer a convex function.

4. Inequalities for stochastic processes. In this section we reformulate the problems into stochastic processes, where σ (or s), which previously was a control variable, is now arbitrary and possibly unknown. We wish to establish some inequalities of the probability of hitting an upper bound u before some given time T and before hitting a lower bound l .

Consider the random walk $X_t = X_{t-1} + \sigma_{t-1} Y_{t-1}$, where Y_{t-1} has the two-point distribution $P(Y = -1) = 1 - w$ and $P(Y = (1 - r)/r) = w$, $w > r$ and where σ_{t-1} is arbitrary. Let F , k , and q be defined as in § 2.1 and let $\tau_u = \min \{t: X_t \geq u\}$ and $\tau_l = \min \{t: X_t \leq l\}$.

INEQUALITY 1. *If $w \geq \frac{1}{2} \geq r$ and $l < X_0 < u$, then*

$$\begin{aligned} P(\tau_u \leq T, \tau_u < \tau_l | X_0) &\leq F(k-1 | T, 1-w) + \left[q(f, T) \binom{T}{k} \right] (1-w)^k w^{T-k} \\ &< F(k | T, 1-w), \end{aligned}$$

where $f = (X_0 - l)/(u - l)$ and $[x]$ is the integer part of x .

Proof. The proof follows from Theorem 3, where we found the strategy σ^* that maximizes the above probability.

INEQUALITY 2. *For any $w > r$ and if $l < X_0 < u$, then*

$$P(\tau_u \leq T, \tau_u < \tau_l | X_0) \leq F(k-1 | T, 1-w) + q(f, T) \binom{T}{k} (1-w)^k w^{T-k} < F(k | T, 1-w).$$

Proof. The proof follows from Corollary 2.

Remark. Since $P(X_T \geq u, X_s > l \text{ for all } s < T | X_0) \leq P(\tau_u \leq T, \tau_u < \tau_l | X_0)$, we get inequalities for this probability, also.

We now turn to the continuous time process $dX_t = \sigma_t(\mu_t dt + dB_t)$, where B_t is the standard Brownian motion and where σ_t is arbitrary but nonnegative and fulfils the conditions stated in the beginning of § 3.

INEQUALITY 3. *If $l < X_0 < u$, then*

$$P(X_T \geq u, X_s > l \forall s < T | X_0) \leq P(\tau_u \leq T, \tau_u < \tau_l | X_0) \leq \Phi\left(\Phi^{-1}\left(\frac{X_0 - l}{u - l}\right) + m_t\right),$$

where m_t is defined as in § 3.

Proof. The proof follows from Theorem 7.

Remark. The inequality also holds for a process $dX_t = \sigma_t(\tilde{\mu}_t dt + dB_t)$, where $\sigma_t \tilde{\mu}_t \leq \sigma_t \mu_t$ for all t .

Acknowledgment. The author thanks David Heath for valuable discussions.

REFERENCES

- [1] L. BREIMAN, *Optimal gambling systems for favorable games*, in Proc. of the 4th Berkeley Sympos. on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, CA, 1961, pp. 65-78.

- [2] L. E. DUBINS AND L. J. SAVAGE, *Inequalities for Stochastic Processes (How to Gamble if You Must)*, Dover, New York, 1965, 1976.
- [3] W. FELLER, *An Introduction to Probability Theory and Its Applications*, John Wiley, New York, 1957.
- [4] D. HEATH, S. OREY, V. PESTIEN, AND W. SUDDERTH, *Minimizing or maximizing the time to reach zero*, SIAM J. Control Optim., 25 (1987), pp. 195–205.
- [5] J. L. KELLY, JR., *A new interpretation of information rate*, Bell System Tech. J., 35 (1956), pp. 917–926.
- [6] W. D. SUDDERTH AND A. WEERASINGHE, *Controlling a process to a goal in finite time*, Math. Oper. Res., 14 (1989), pp. 400–409.
- [7] B. W. TURNBULL, *Chebyshev-like inequalities for dam models*, J. Appl. Probab., 9 (1972), pp. 617–629.

ON A CERTAIN ASYMPTOTIC BEHAVIOUR OF TIME-INVARIANT RICCATI MATRIX DIFFERENTIAL EQUATIONS*

WERNER KRATZ†

Abstract. The main result of this paper reads as follows: Given (real) $(n \times n)$ -matrices A , B , C , and C_0 such that B , C , and C_0 are symmetric, such that B and C_0 are nonnegative definite, such that the pair (A, B) is (completely) controllable (i.e., $\text{rank} [B, A, \dots, A^{n-1}B] = n$), and such that the triple (A, B, C_0) is strongly observable (i.e., $\dot{x} = Ax + Bu$, $C_0 x \equiv 0$ on some nondegenerate interval implies $x(t) \equiv 0$), then, for any $t_0 > 0$ and any symmetric matrix Q_0 , the solution $Q(t; \lambda)$ of the Riccati matrix differential equation

$$\dot{Q} + A^T Q + Q A + Q B Q - C + \lambda C_0 = 0, \quad Q(0) = Q_0$$

exists on $[0, t_0]$ if $\lambda \leq \lambda_0$, and it satisfies $\lim_{\lambda \rightarrow -\infty} Q(t_0; \lambda) = \infty$ (i.e., all eigenvalues of the symmetric matrix Q tend to ∞). The result is that the new notion of strong observability is even necessary for the assertion. This result on Riccati matrix equations is motivated by the following application. It is shown that

$$\min \left\{ \int_0^{t_0} [x^T C x + u^T B u] dt \middle/ \int_0^{t_0} x^T C_0 x dt, \text{ where } x(t) \neq 0, \text{ and } \dot{x} = Ax + Bu \text{ with } u \in C^s[0, t_0] \right\}$$

exists just under the same assumptions as above.

Key words. Riccati matrix differential equation, linear systems, controllability, observability, optimal linear regulator, Rayleigh quotient

AMS(MOS) subject classifications. 49A10, 34A10, 93B05, 93B07, 93C45

1. Introduction. The main objective of this paper is a result on the asymptotic behaviour of certain Riccati matrix differential equations, which is discussed below. These investigations, however, were motivated mainly by the following optimization problem. Given a time-invariant linear system

$$\dot{x} = Ax + Bu, \quad y = C_0 x$$

with state x , input u , and output y , we ask whether the

$$(*) \quad \min \{R(x) \mid \dot{x} = Ax + Bu \text{ for some } u \in C^s[0, t_0], x \neq 0\}$$

exists, where the “Rayleigh quotient” $R(x)$ is given by

$$R(x) = \int_0^{t_0} \{x^T(t) C x(t) + u^T(t) B u(t)\} dt \middle/ \int_0^{t_0} x^T(t) C_0 x(t) dt,$$

where $C^s[0, t_0]$ denotes the *piecewise continuous* functions on $[0, t_0]$, where we assume naturally that C , B , C_0 are symmetric $(n \times n)$ -matrices, and where B and C_0 are nonnegative definite (observe that the nonnegative-definiteness of B corresponds to the Legendre condition in the calculus of variations). In this paper, we give conditions on the matrices involved that guarantee the existence of the minimum above. Of course, this problem is closely related to the linear optimal regulator, where only the quadratic functional in the numerator of $R(x)$ is considered, and its theory is well developed. Here, we seek to minimize the numerator relative to another quadratic functional (which can be interpreted as output-performance of the linear system). Moreover, the question of minimizing $R(x)$ is a classical problem in the calculus of variations (here $R(x)$ is the so-called Rayleigh quotient of corresponding eigenvalue problems; see,

* Received by the editors March 20, 1991; accepted for publication (in revised form) August 13, 1991.

† Abteilung Mathematik, Universität Ulm, Albert-Einstein-Allee 11, D-7900 Ulm, Germany.

e.g., [21–23]), and it is completely solved only for special cases, namely, quadratic functionals (where B and C_0 are regular) and for so-called Sturm–Liouville problems (where $\text{rank } B = \text{rank } C_0 = 1$) [3], [21]–[23]. Except in these two cases, there was always assumed some definiteness property (which reduces to the assumption that $R(x)$ is bounded from below; see, e.g., [23, p. 381, (\mathcal{H}_κ)], [4], [11], [18], [21], [22], [25], and others]) to prove that the above minimum exists (which corresponds to the existence of a (minimal) eigenvalue of the related eigenvalue problem).

As we see below, the problem of minimizing $R(x)$ is intimately connected with the asymptotic behaviour of a corresponding matrix differential equation, which plays also an important role for the linear optimal regulator [2], [7], [10], [17]. Our main result (Theorem 3, below) reads as follows:

Given (real) $(n \times n)$ -matrices A , B , C , and C_0 such that B , C , and C_0 are symmetric, such that B and C_0 are nonnegative definite, such that the pair (A, B) is *controllable* (i.e., $\text{rank } [B, A, \dots, A^{n-1}B] = n$), and such that the triple (A, B, C_0) is *strongly observable* (i.e., $\dot{x} = Ax + Bu$, $C_0 x \equiv 0$ on some nondegenerate interval implies that $x \equiv 0$). Then, for any $t_0 > 0$ and any symmetric matrix Q_0 , the solution $Q(t; \lambda)$ of the Riccati matrix differential equation

$$(\quad +) \quad \dot{Q} + A^T Q + QA + QBQ - C + \lambda C_0 = 0, \quad Q(0) = Q_0$$

exists on $[0, t_0]$ if $\lambda \leq \lambda_0$, and it satisfies

$$\lim_{\lambda \rightarrow -\infty} Q(t_0; \lambda) = \infty$$

(i.e., all eigenvalues of the symmetric matrix Q tend to ∞).

Using this result, we prove in Theorem 4 that the minimum (*) does exist just under the same assumptions. In Proposition 6, we show that the new notion of strong observability is also necessary for the assertion of Theorem 3 (i.e., the asymptotic behaviour of $Q(t; \lambda)$, above). This asymptotic behaviour is known (even in a more explicit form) when we deal with special cases coming from Sturm–Liouville problems [14], [15], [20] or from quadratic functionals [3], as mentioned above. Throughout this paper, we restrict ourselves to constant matrices, but inequalities related to Riccati equations (see, e.g., [24]) lead to corresponding results for time-dependent systems.

Now we summarize the setup of this paper. In § 2 we derive a normal form for controllable systems, which is related to the standard normal form used to show the theorem on pole assignment (see, e.g., [12], [19], [26]). In § 3 we prove a central inequality (Theorem 2), which is the key for the proof of our main result. This inequality (together with the normalization of § 2) leads to a form of C_0 , such that the general Riccati equation (+) separates into special equations, which are related to Sturm–Liouville problems where the asymptotic result is known (see Proposition 5, below). In § 4 we state the correspondence of (+) to certain linear systems in the sense that Q solves (+) if and only if

$$Q = UX^{-1}, \quad \text{where } \dot{X} = AX + BU, \quad \dot{U} = (C - \lambda C_0)X - A^T U,$$

and this is used to derive inequalities on Riccati equations (Proposition 3). While § 5 is devoted to the proof of our main result (Theorem 3), we apply this result in § 6 to show the existence of the minimum (*) (Theorem 4). This proof of Theorem 4 uses a standard technique known from the linear optimal regulator. Moreover, we prove the necessity of strong observability (Proposition 6) in § 6, as discussed above.

2. The normal form. In this section, we derive a canonical form of (completely) controllable systems, which is the basis for the main theorem in § 5. The first part

(Proposition 1) of this normal form is contained in [13, Prop. 1], where the transformation is constructed explicitly by induction], and it essentially equals the so-called Hessenberg form [26, eq. (5.186)]. Moreover, our final canonical form (Theorem 1) can be used to very easily derive the standard normal form for proving the theorem on pole assignment [1], [19], [27], [12, Satz 3.5 "Regelungsnormalform"] (see also [6], [9]). Hence our normal form (and its proof) can be used to give a rather simple alternative proof of pole assignment for controllable systems (either using Theorem 1 directly or by deriving first the standard normal form as in [1] or [12], [19]).

PROPOSITION 1. Assume that the pair (A, B) of (real) $(n \times n)$ -matrices A and B is controllable, i.e., $\text{rank}[B, AB, \dots, A^{r-1}B] = n$ for some (minimal) $r \in \{1, \dots, n\}$, and let $l_r = \text{rank } B$, $\sum_{\mu=r-k+1}^r l_\mu = \text{rank}[B, AB, \dots, A^{k-1}B]$ for $k=1, \dots, r$. Then $1 \leq l_1 \leq \dots \leq l_r$, $\sum_{\mu=1}^r l_\mu = n$, and then there exists an orthogonal matrix T such that

$$(i) \quad T^T B = \begin{pmatrix} 0 \\ -\frac{0}{B_0} \end{pmatrix}$$

with an $(l_r \times n)$ -matrix B_0 of rank l_r ; and

(ii) $T^T A T = (A_{\mu\nu})$, $\mu, \nu \in \{1, \dots, r\}$ with $(l_\mu \times l_\nu)$ -matrices $A_{\mu\nu}$, where $A_{\mu\nu} = 0$ for $\nu > \mu + 1$ and $\text{rank } A_\mu = l_\mu$ for $A_\mu = A_{\mu, \mu+1} \dots A_{\mu-1, r}$, $\mu = 1, \dots, r-1$ (see [13, Prop. 1], and observe that r is the so-called controllability index of the system $\dot{x} = Ax + Bu$, according to [19]).

Now we can derive our normal form.

THEOREM 1. Under the assumptions and with the notation of Proposition 1, there exists a regular matrix T such that

$$(i) \quad T^{-1} B = \begin{pmatrix} 0 \\ -\frac{0}{B_0} \end{pmatrix}$$

with an $(l_r \times n)$ -matrix B_0 of rank l_r as in Proposition 1; and

$$(ii) \quad T^{-1} A T = \begin{pmatrix} 0 & J_1 & \dots & 0 \\ & \ddots & & \vdots \\ 0 & & 0 & J_{r-1} \\ A_{r1} & \dots & & A_{rr} \end{pmatrix}$$

with $(l_\mu \times l_{\mu+1})$ matrices J_μ of the form $J_\mu = (I \mid 0)$, where $I = I_\mu$ is the $(l_\mu \times l_\mu)$ -identity matrix, and with certain $(l_r \times l_\mu)$ -matrices $A_{r\mu}$.

Proof. We may assume that

$$A = (A_{\mu\nu}^{(0)}), \quad B = \begin{pmatrix} 0 \\ -\frac{0}{B_0} \end{pmatrix}$$

are as in the statement of Proposition 1. First, let $T = \text{diag}(T_1, \dots, T_r)$ be a block-diagonal matrix with regular $(l_\mu \times l_\mu)$ -matrices T_μ and with $T_1 = I$. Then $T^{-1} A T = A^{(1)} = (A_{\mu\nu}^{(1)})$ with $A_{\mu\nu}^{(1)} = T_\mu^{-1} A_{\mu\nu}^{(0)} T_\nu$. Hence $A_{\mu\nu}^{(1)} = 0$ for $\nu > \mu + 1$, and, inductively, $J_\mu = A_{\mu, \mu+1}^{(1)} = T_\mu^{-1} A_{\mu, \mu+1} T_{\mu+1} = (I \mid 0)$ for $\mu = 1, \dots, r-1$ with suitable, regular matrices T_2, \dots, T_r , since $\text{rank } A_{\mu, \mu+1} = l_\mu$. Moreover, the form of B remains unchanged when multiplying with T^{-1} from the left. Next, let

$$T = \begin{pmatrix} I & 0 & 0 \\ \left(-A_{11}^{(1)} \right) & I & \vdots \\ 0 & & \ddots \\ \vdots & & & \vdots \\ 0 & \dots & & I \end{pmatrix}.$$

Then

$$T^{-1} = \begin{pmatrix} I & 0 & 0 \\ \begin{pmatrix} A_{11}^{(1)} \\ 0 \end{pmatrix} & I & \vdots \\ \vdots & & \ddots \\ 0 & \dots & I \end{pmatrix}$$

and

$$A^{(2)} = T^{-1}A^{(1)}T = \begin{pmatrix} 0 & J_1 & \dots & 0 \\ A_{21}^{(2)} & A_{22}^{(2)} & J_2 & 0 \\ & & \ddots & \\ A_{r1}^{(2)} & & & A_{rr}^{(2)} \end{pmatrix}.$$

After r steps these transformations lead to the final form, namely,

$$A^{(r)} = \begin{pmatrix} 0 & J_1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & J_{r-1} \\ A_{r1}^{(r)} & \dots & & A_{rr}^{(r)} \end{pmatrix},$$

which is asserted in our theorem, while B remains unchanged. \square

In view of our application of this normal form, we can simplify this form further by dealing with inequalities instead of identities. This is motivated by the monotonicity behaviour of Riccati equations (as discussed below). We henceforth write $A \leq B$ for square-matrices A, B , if A and B are symmetric and if $B - A$ is nonnegative definite.

COROLLARY 1. *Under the assumptions and with the notation of Proposition 1 (respectively Theorem 1), suppose, moreover, that $B \geq 0$ and that C is another $(n \times n)$ -matrix, which is symmetric. Then there exists a regular matrix T and a real $\alpha > 0$ such that*

$$\begin{pmatrix} \tilde{C} & -\tilde{A}^T \\ -\tilde{A} & -\tilde{B} \end{pmatrix} \geq \begin{pmatrix} -\alpha I & -J^T \\ -J & -\mathcal{B} \end{pmatrix},$$

where $\tilde{A} = T^{-1}AT$, $\tilde{B} = T^{-1}B(T^{-1})^T$, $\tilde{C} = T^TCT$,

$$J = \begin{pmatrix} 0 & J_1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & J_{r-1} \\ 0 & \dots & & 0 \end{pmatrix}$$

with $J_\mu = (I_\mu | 0)$ for $\mu = 1, \dots, r-1$ as in Theorem 1, and where $\mathcal{B} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$ with $(l_r \times l_r)$ -identity matrix I .

Proof. Suppose, first, that T is a regular matrix as in the statement of Theorem 1. Then

$$T^{-1}B(T^{-1})^T = \begin{pmatrix} 0 & 0 \\ 0 & B_0 \end{pmatrix}$$

with a positive definite $(l_r \times l_r)$ -matrix B_0 , since B is symmetric, nonnegative definite, and $\text{rank } B = \text{rank } B_0 = l_r$. Moreover, we may assume that $2B_0 \leq I$ (multiply T by some positive number, which does not change \tilde{A}). Now, given any vectors $c, d \in \mathbb{R}^n$, let

$c^T = (c_1^T, \dots, c_r^T)$, $d^T = (d_1^T, \dots, d_r^T)$ with $c_\nu, d_\nu \in \mathbb{R}^{l_\nu}$. Then

$$\begin{aligned} & c^T(\tilde{C} + \alpha I)c + d^T(\mathcal{B} - \tilde{B})d + 2d^T(J - \tilde{A})c \\ &= c^T(\tilde{C} + \alpha I)c + d_r^T(I - B_0)d_r - 2d_r^T \sum_{\nu=1}^r A_{r\nu}c_\nu \\ &\geq \frac{1}{2} \left| d_r - 2 \sum_{\nu=1}^r A_{r\nu}c_\nu \right|^2 + c^T(\tilde{C} + \alpha I - 2(\tilde{A} - J)(\tilde{A} - J)^T)c \geq 0, \end{aligned}$$

if $\alpha > 0$ is sufficiently large. \square

Next, we give those transformations T , under which the normal form of Corollary 1 is (essentially) invariant. To do this, we must refine the block-structure of A (respectively, J) as follows: Let l_1, \dots, l_r be given as before (i.e., as in Theorem 1 and elsewhere) such that $1 \leq l_1 \leq \dots \leq l_r$, $\sum_{\nu=1}^r l_\nu = n$. We subdivide the blocks of length l_ν , further, namely, into ν blocks of lengths $\Delta_\mu = l_\mu - l_{\mu-1} \geq 0$ (with $l_0 := 0$). Then $\sum_{\mu=1}^\nu \Delta_\mu = l_\nu$ for $\nu = 1, \dots, r$. Using this notation, we obtain the following result.

COROLLARY 2. Assume that J and \mathcal{B} are given as in Corollary 1, and let $\Delta_\mu = l_\mu - l_{\mu-1}$ for $\mu = 1, \dots, r$, $l_0 := 0$. Moreover, let $\nu \in \{1, \dots, r\}$, and suppose that T is a regular matrix of the following form: $T^{-1} = \text{diag}(T_1, \dots, T_r)$ is block-diagonal, where the blocks T_μ along the diagonal are $(l_\mu \times l_\mu)$ -matrices such that $T_\mu = I$ for $\mu = 1, \dots, \nu-1$,

$$T_\nu = \begin{pmatrix} I & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & & I & 0 \\ C_{\nu 1} & \dots & & C_{\nu \nu} \end{pmatrix}$$

with arbitrary $(\Delta_\nu \times \Delta_\mu)$ -matrices $C_{\nu\mu}$, where $C_{\nu\nu}$ is regular, and let

$$T_{\nu+1} = \begin{pmatrix} I & \dots & 0 \\ \vdots & \ddots & \vdots \\ C_{\nu 1} & \dots & C_{\nu \nu} \\ 0 & \dots & I \end{pmatrix}, \dots, T_r = \begin{pmatrix} I & \dots & 0 \\ \vdots & \ddots & \vdots \\ C_{\nu 1} & \dots & C_{\nu \nu} \\ 0 & \dots & I \end{pmatrix} \leftarrow \nu\text{th row.}$$

Then $T^{-1}JT = J$, $T^{-1}\mathcal{B}(T^{-1})^T = \begin{pmatrix} 0 & 0 \\ 0 & B_0 \end{pmatrix}$ with a positive definite $(l_r \times l_r)$ -matrix B_0 ($B_0 = T_r^T T_r$), and, of course, $T^T T \leq \alpha I$ for suitable $\alpha > 0$.

Proof. The statements of the corollary follow quite easily if we show that $T_\mu J_\mu T_{\mu+1}^{-1} = J_\mu$ for $\mu = 1, \dots, r-1$. This is trivial for $\mu = 1, \dots, \nu-2$, and, for $\mu = \nu-1$, we have that

$$T_{\nu-1} J_{\nu-1} T_\nu^{-1} = \left(I \begin{smallmatrix} 1 \\ \vdots \\ 1 \end{smallmatrix} \begin{smallmatrix} 0 \\ \vdots \\ 0 \end{smallmatrix} \right) \begin{pmatrix} I & 0 \\ * & * \end{pmatrix} = \left(I \begin{smallmatrix} 1 \\ \vdots \\ 1 \end{smallmatrix} \begin{smallmatrix} 0 \\ \vdots \\ 0 \end{smallmatrix} \right) = J_{\nu-1}.$$

Finally, for $\mu = \nu, \dots, r-1$, we obtain that

$$T_\mu J_\mu T_{\mu+1}^{-1} = T_\mu \left(I \begin{smallmatrix} 1 \\ \vdots \\ 1 \end{smallmatrix} \begin{smallmatrix} 0 \\ \vdots \\ 0 \end{smallmatrix} \right) \begin{pmatrix} T_\mu^{-1} & 0 \\ 0 & I \end{pmatrix} = J_\mu. \quad \square$$

DEFINITION 1. Transformations T , which are a (finite) product of matrices as described in Corollary 2 above, are called *invariant matrices*.

3. The central inequality. The result of this section is the key to the proof of our theorem on the asymptotic behaviour of Riccati equations. Before we can state the

inequality, however, we need some notion. Let (as in the preceding section)

$$\begin{aligned} n \in \mathbb{N}, \quad 1 \leq l_1 \leq \cdots \leq l_r \quad \text{with} \quad \sum_{\nu=1}^r l_\nu = n, \quad \text{and} \\ \Delta_\nu = l_\nu - l_{\nu-1}, \quad \nu = 1, \dots, r \quad \text{with} \quad l_0 := 0, \quad \text{such that} \\ l_\nu = \sum_{\mu=1}^{\nu} \Delta_\mu, \quad \text{in particular} \quad m := \sum_{\mu=1}^r \Delta_\mu = l_r; \end{aligned}$$

the $(n \times n)$ -matrices J and \mathcal{B} are defined as in Corollary 1. Then a function $x \in C_1(\mathcal{J})$ (on some interval $\mathcal{J} \subset \mathbb{R}$) satisfies

$$\dot{x} = Jx + \mathcal{B}u \quad \text{on } \mathcal{J} \text{ for some function } u(t) \in C^s(\mathcal{J})$$

if and only if $x(t)$ is of the following form:

$$(1) \quad \begin{aligned} x^T(t) &= (x_1^T(t), x_1''^T(t), x_2^T(t), x_1''^T(t), x_2'^T(t), x_3^T(t), \dots, x_r^T(t)) \\ &\quad \text{with } x_\nu(t) \in \mathbb{R}^{\Delta_\nu}, \in C_{r-\nu+1}^s(\mathcal{J}) \text{ for } \nu = 1, \dots, r. \end{aligned}$$

(By $x, x', x'', x^{(\nu)}$, and so forth, we denote the first, second, ν th, and other derivatives of x .)

DEFINITION 2. Functions $x(t)$ of the form (1) are called *admissible* on \mathcal{J} , and, for such functions, we define

$$\tilde{x}^T(t) := (x_1^T(t), \dots, x_r^T(t)) = (y_1(t), y_2(t), \dots, y_m(t)) \in \mathbb{R}^m.$$

Observe that $y_1(t), \dots, y_m(t)$ are just the “free” real-valued functions defining the corresponding admissible $x(t)$. Moreover, we want to mention that we allow, in particular, $\Delta_\nu = 0$ for $\nu \in \{2, \dots, r\}$, such that the corresponding $x_\nu(t)$ does not occur.

THEOREM 2. Let C be a symmetric and nonnegative definite $(n \times n)$ -matrix and $t_0 > 0$, such that (compare Definition 3 below)

$$f(x) := \int_0^{t_0} x^T(t) C x(t) dt > 0$$

for all admissible functions $x(t)$ on $[0, t_0]$, which do not vanish identically. Then there exist constants $\delta \geq 1$, $\varepsilon_0 > 0$ and an invariant transformation T such that the following holds:

$$(2) \quad \begin{aligned} f(Tx) &\geq \int_0^{t_0} \left\{ \sum_{\nu=1}^r \sum_{\mu=l_{\nu-1}+1}^{l_\nu} [\varepsilon_\mu |y_\mu|^2 - \delta \varepsilon_{\mu+1} (|y'_\mu|^2 + \dots + |y_\mu^{(r-\nu)}|^2)] \right\} dt \\ &= \int_0^{t_0} \{ \varepsilon_1 |y_1|^2 - \delta \varepsilon_2 (|y'_1|^2 + \dots + |y_1^{(r-1)}|^2) + \dots + \varepsilon_m |y_m|^2 \} dt \end{aligned}$$

for all admissible functions $x(t)$ on $[0, t_0]$ and for all $\varepsilon_1, \dots, \varepsilon_m$ with $0 < \varepsilon_\nu \leq \varepsilon_0$, $\delta \cdot \varepsilon_{\nu+1} \leq \varepsilon_\nu$ for $\nu = 1, \dots, m$ ($\varepsilon_{m+1} := 0$).

Proof. We proceed by induction with respect to $n \in \mathbb{N}$. First, if $n = 1$, then $r = m = \Delta_1 = 1$, and the assumption on $f(x)$ implies that $C > 0$. Hence

$$f(x) = \int_0^{t_0} C x^2(t) dt \geq \varepsilon_1 \int_0^{t_0} x^2(t) dt \quad \text{if } 0 < \varepsilon_1 \leq \varepsilon_0 := C,$$

which yields the induction hypothesis for $n = 1$. Next, let $n \geq 2$, and suppose throughout that $x(t)$ is admissible. We consider two cases.

Case 1. We have that $\Delta_r \geq 1$. Let $x_r^T(t) = (\bar{x}_r^T(t), y(t))$ with real-valued $y(t)$ (i.e., $\bar{x}_r(t)$ denotes the first $\Delta_r - 1$ coordinates of $x_r(t)$, and $\bar{x}_r^T(t)$ denotes its transpose).

Then $x^T(t) = (\bar{x}^T(t), y(t))$, where $\bar{x}(t)$ is admissible according to Definition 2, but $\bar{x}(t) \in \mathbb{R}^{n-1}$ instead of \mathbb{R}^n with corresponding “parameters” $\bar{l}_\nu = l_\nu$, $\bar{\Delta}_\nu = \Delta_\nu$ for $\nu = 1, \dots, r-1$ and $\bar{l}_r = l_r - 1$, $\bar{\Delta}_r = \Delta_r - 1 \geq 0$, $\bar{m} = m - 1$ such that the induction hypothesis may be applied to these $\bar{x}(t)$. We write

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

such that C_{11} is an $((n-1) \times (n-1))$ -matrix, $C_{21}^T = C_{12} \in \mathbb{R}^n$, and such that $C_{22} \in \mathbb{R}$. Then

$$x^T C x = \bar{x}^T C_{11} \bar{x} + 2 \bar{x}^T C_{12} y + C_{22} y^2.$$

Here we see the main assumption, i.e., $f(x) = \int_0^{t_0} x^T C x dt > 0$ if $x(t) \neq 0$ (compare the notion of strong observability in Definition 3). It implies that $C_{22} > 0$, since otherwise we would have that $f(x) = 0$ (i.e., $x^T C x \equiv 0$ on $[0, t_0]$) for $\bar{x}_r(t) \equiv 0$ and arbitrary, nonvanishing $y(t) \in C^s[0, t_0]$, which yields an admissible $x(t) \neq 0$ by Definition 2. Hence

$$(3) \quad x^T C x = \bar{x}^T \bar{C} \bar{x} + C_{22} \left(y + \frac{1}{C_{22}} C_{12}^T \bar{x} \right)^2$$

with $\bar{C} = C_{11} - (1/C_{22}) C_{12} C_{12}^T$. Then \bar{C} is nonnegative definite (since C is nonnegative definite), and $\bar{f}(\bar{x}) := \int_0^{t_0} \bar{x}^T \bar{C} \bar{x}(t) dt > 0$ for all admissible $\bar{x}(t) \neq 0$ (put $y(t) = -(1/C_{22}) C_{12}^T \bar{x}(t)$ and use (3)). Thus we may apply the induction hypothesis, i.e., that there exist $\bar{\delta} \geq 1$, $\bar{\varepsilon}_0 > 0$, and in invariant transformation $\bar{T}_{(n-1) \times (n-1)}$ such that

$$(4) \quad \bar{f}(\bar{T} \bar{x}) \geq \int_0^{t_0} \left\{ \sum_{\nu=1}^r \sum_{\mu=\bar{l}_{\nu-1}+1}^{\bar{l}_\nu} [\varepsilon_\mu |\bar{y}_\mu|^2 - \bar{\delta} \varepsilon_{\mu+1} (|\bar{y}'_\mu|^2 + \dots + |\bar{y}_\mu^{(r-\nu)}|^2)] \right\} dt$$

for all admissible $\bar{x}(t)$ and all $\varepsilon_1, \dots, \varepsilon_{m-1}$, $\varepsilon_m := 0$ with $0 < \varepsilon_\nu \leq \bar{\varepsilon}_0$, $\bar{\delta} \varepsilon_{\nu+1} \leq \varepsilon_\nu$ for $\nu = 1, \dots, m-1$. Now, for $0 < \varepsilon_m \leq \frac{1}{2} C_{22}$, we obtain that

$$\begin{aligned} C_{22} \left(y + \frac{1}{C_{22}} C_{12}^T \bar{T} \bar{x} \right)^2 &= \varepsilon_m y^2 + (C_{22} - \varepsilon_m) \left(y + \frac{1}{C_{22} - \varepsilon_m} C_{12}^T \bar{T} \bar{x} \right)^2 \\ &\quad - \frac{\varepsilon_m}{C_{22}(C_{22} - \varepsilon_m)} (C_{12}^T \bar{T} \bar{x})^2 \geq \varepsilon_m y^2 - \delta^* \varepsilon_m |\bar{x}|^2 \\ &= \varepsilon_m y^2 - \delta^* \varepsilon_m \sum_{\nu=1}^r \sum_{\mu=\bar{l}_{\nu-1}+1}^{\bar{l}_\nu} [|\bar{y}_\mu|^2 + \dots + |\bar{y}_\mu^{(r-\nu)}|^2] \end{aligned}$$

if $\delta^* = \delta^*(C_{22}, C_{12}^T, \bar{T})$ is sufficiently large. Using (3) and (4), we obtain the induction hypothesis for some $\bar{\delta} \geq 2\bar{\delta} \geq 1$, $0 < \varepsilon_0 \leq \bar{\varepsilon}_0$ with

$$T = \begin{pmatrix} \bar{T} & 0 \\ 0 & 1 \end{pmatrix}_{n \times n}$$

(which is an invariant transformation, too).

Case 2. We have that $\Delta_r = 0$. Then $r \geq 2$, and the admissible functions $x(t)$ are of the form $x^T(t) = (x_1^T(t), x_1'^T(t), \dots, x_1^{(r-1)T}(t), \dots, x_{r-1}^T(t))$, so that $l := n - m \geq m$. We write

$$C = \left(\underbrace{\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}}_l \right)_l^l \left(\underbrace{\quad}_m \right)_m^m,$$

and, according to this notion, we split this into two subcases.

Case 2.1. C_{22} is singular. This is the main case, where the invariant transformations of Corollary 2 are needed; actually, the whole degree of freedom of our normal form is exhausted. Since C_{22} is singular and symmetric, there exists a lower-triangular and regular $(m \times m)$ -matrix T_r and some $\rho \in \{1, \dots, m\}$ such that the ρ th column and the ρ th row of $\tilde{C}_{22} := (T_r^{-1})^T C_{22} T_r^{-1}$ are zero. Such a matrix T_r is obviously a product of matrices T_r , as occurring in Corollary 2; i.e.,

$$T_r = \begin{pmatrix} C_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ C_{r1} & \cdots & C_{rr} \end{pmatrix} \quad \text{with } (\Delta_\nu \times \Delta_\mu)\text{-matrices } C_{\nu\mu}.$$

Moreover, by Corollary 2, if $T^{-1} = \text{diag}(T_1, \dots, T_r)$ with

$$T_\nu = \begin{pmatrix} C_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ C_{\nu 1} & \cdots & C_{\nu \nu} \end{pmatrix} \quad \text{for } \nu = 1, \dots, r,$$

then T is an invariant matrix. We have that $f(Tx) = \int_0^{t_0} x^T(t) \bar{C}x(t) dt$, where

$$\bar{C} = T^T C T = \begin{pmatrix} \bar{C}_{11} & \bar{C}_{12} \\ \bar{C}_{21} & \bar{C}_{22} \end{pmatrix}.$$

Since the ρ th column and the ρ th row of \bar{C}_{22} are zero, and since C is nonnegative definite, it follows that the ρ th column of \bar{C}_{12} , respectively, the ρ th row of $\bar{C}_{21} = \bar{C}_{12}^T$, is zero (i.e., the whole corresponding column, respectively row, of \bar{C} is zero). This means that the highest derivative of the function y_ρ (where $\hat{x}^T = (y_1, \dots, y_m)$ as in Definition 2) does not occur in the quadratic form $x^T \bar{C} x$ at all. Hence the induction hypothesis for $n-1$ yields the assertion for n .

Case 2.2. C_{22} is regular. In this case, we prove a stronger assertion, namely, that there exists $\alpha > 0$ such that

$$(5) \quad f(x) \geq \alpha \int_0^{t_0} \{|y_1|^2 + \cdots + |y_m|^2\} dt$$

for all admissible $x(t)$.

In this case (i.e., $\Delta_r=0$ with regular C_{22}), our assertion can be reduced to a situation, where Corollary 3 to Theorem 4 below is applicable, as follows. For any admissible $x(t)$, write $x^T(t) = (\bar{x}^T(t), u^{*T}(t))$ with $\bar{x} \in \mathbb{R}^l$ and $u^* \in \mathbb{R}^m$ such that $u^* = (x_1^{(r-1)T}, \dots, x_{r-1}^{(r-1)T})$. Moreover, let

$$\begin{aligned} \bar{A}_{l \times l} &:= \left(\begin{array}{cccc} 0 & J_1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & J_{r-2} \\ & -C_{22}^{-1} & C_{21} & \end{array} \right) \left. \vphantom{\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}} \right\}^{l-m}_m, \quad \bar{B}_{l \times l} := \left(\begin{array}{cc} 0 & 0 \\ 0 & C_{22}^{-1} \end{array} \right) \left. \vphantom{\begin{pmatrix} 0 \\ 0 \end{pmatrix}} \right\}^{l-m}_m, \\ \bar{C}_{l \times l} &:= C_{11} - C_{21}^T C_{22}^{-1} C_{21}, \quad \bar{u}(t) := \begin{pmatrix} 0 \\ C_{22} u^*(t) + C_{21} \bar{x}(t) \end{pmatrix} \in \mathbb{R}^l. \end{aligned}$$

Then the pair (\bar{A}, \bar{B}) is controllable (even in normal form as in Theorem 1), $\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}\bar{u}$ (if and only if x is admissible), and $x^T C x = \bar{x}^T \bar{C}_1 \bar{x} + \bar{u}^T \bar{B} \bar{u}$. It follows that $f(x) = \int_0^t \{\bar{x}^T \bar{C}_1 \bar{x} + \bar{u}^T \bar{B} \bar{u}\} dt > 0$ for all admissible x , i.e., $\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}\bar{u}$ with $x \neq 0$. Now Corollary 3 below for dimension $l < n$ implies that there exists $\alpha > 0$ (depending on $\bar{C}_1, \bar{A}, \bar{B}$,

and also on t_0) such that

$$f(x) \cong \alpha \int_0^{t_0} |\bar{x}(t)|^2 dt \cong \alpha \int_0^{t_0} \{|y_1|^2 + \cdots + |y_m|^2\} dt$$

for all admissible $x(t)$ (choose the identity matrix as corresponding $l \times l$ -matrix C_0). This completes the proof of Theorem 2. \square

Remarks. (i) The foregoing proof shows that the matrix T and the constant $\delta \geq 1$ depend on the matrix C only, while the constant $\varepsilon_0 > 0$ depends additionally on t_0 (because of Case 2.2., where Corollary 3 to Theorem 4 is applied; observe the corresponding remark there).

(ii) Note, that $\Delta_\nu = 0$ for some $\nu \in \{2, \dots, r\}$ is explicitly allowed (while $\Delta_1 = l_1 \geq 1$). Then $l_{\nu-1} = l_\nu$, and $\sum_{\mu=l_{\nu-1}+1}^{l_\nu} \cdots$ is empty, i.e., $= 0$. Moreover, the sum $|y'_\mu|^2 + \cdots + |y^{(r-\nu)}_\mu|^2$ is also $= 0$, if $\nu = r$.

4. Auxiliary results on Riccati matrix differential equations. We need some statements on Riccati equations, which are essentially known, so we omit the proofs, except for Proposition 3, which differs substantially from the cited results. Throughout, we assume that A, B, C , and Q_0 are any (real) $(n \times n)$ -matrices, and we consider the initial value problem for the Riccati matrix differential equation

$$(6) \quad \dot{Q} + A^T Q + Q A + Q B Q - C = 0, \quad Q(t_0) = Q_0$$

on some interval \mathcal{I} of the real axis with $t_0 \in \mathcal{I}$. First, we have the well-known correspondence with a linear system, a so-called *Hamiltonian system* (see, e.g., [15], [16], [23], [24]).

PROPOSITION 2. *An $(n \times n)$ -matrix function $Q(t)$ is a solution of (6) on \mathcal{I} if and only if the following holds: $Q(t) = U(t)X^{-1}(t)$, where the $(n \times n)$ -matrices $X(t)$ and $U(t)$ solve the initial value problem*

$$(7) \quad \dot{X} = AX + BU, \quad \dot{U} = CX - A^T U, \quad X(t_0) = I, \quad U(t_0) = Q_0,$$

and where $X(t)$ is regular for $t \in \mathcal{I}$.

Next, we state the important comparison result, which is essentially contained in [15, Prop. 2] (see also [24, p. 118]).

PROPOSITION 3. *Assume that B, C , and Q_0 are symmetric, that $B \geq 0$ (i.e., nonnegative definite), and that $\mathcal{I} = [t_0, T)$ ($T > t_0$). Moreover, let $\tilde{A}, \tilde{B}, \tilde{C}$, and \tilde{Q}_0 be $(n \times n)$ -matrices that satisfy the fact that \tilde{B}, \tilde{C} , and \tilde{Q}_0 are symmetric such that $\tilde{Q}_0 \leq Q_0$ and such that*

$$(8) \quad \begin{pmatrix} C & -A^T \\ -A & -B \end{pmatrix} \geq \begin{pmatrix} \tilde{C} & -\tilde{A}^T \\ -\tilde{A} & -\tilde{B} \end{pmatrix}.$$

Then, if the solution $\tilde{Q}(t)$ of $\dot{\tilde{Q}} + \tilde{A}^T \tilde{Q} + \tilde{Q} \tilde{A} + \tilde{Q} \tilde{B} \tilde{Q} - \tilde{C} = 0$, $\tilde{Q}(t_0) = \tilde{Q}_0$ exists on \mathcal{I} , the solution $Q(t)$ of (6) exists on \mathcal{I} also, and it satisfies

$$(9) \quad Q(t) \geq \tilde{Q}(t) \text{ for } t \in \mathcal{I} = [t_0, T).$$

Proof. The proof is based on the correspondence to linear systems according to the previous Proposition 2 (so that $Q = UX^{-1}$, respectively, $\tilde{Q} = \tilde{U}\tilde{X}^{-1}$) and on the following identity, which follows from (7) directly (see [15, Prop. 2])

$$\begin{aligned} \frac{d}{dt} X^T (Q - \tilde{Q}) X &= (\tilde{Q} X - U)^T B (\tilde{Q} X - U) \\ &+ (X^T, X^T \tilde{Q}) \begin{pmatrix} C - \tilde{C} & \tilde{A}^T - A^T \\ \tilde{A} - A & \tilde{B} - B \end{pmatrix} \begin{pmatrix} X \\ \tilde{Q} X \end{pmatrix} \geq 0 \end{aligned}$$

since $B \geq 0$, and since (8) holds for all $t \in \mathcal{J}$, where $Q(t)$ exists (observe that $Q(t)$ and $\tilde{Q}(t)$ are symmetric for all t because Q_0 and \tilde{Q}_0 are symmetric). This inequality yields (9) using $Q_0 \geq \tilde{Q}_0$. To prove that $Q(t)$ exists on $[t_0, T)$ we proceed similarly to the proof of Proposition 4 in [15]. If $Q(t)$ does not exist on $[t_0, T)$ then there exists $t_1 \in (t_0, T)$ such that the following holds (use Proposition 2, above): $Q(t)$ exists on $[t_0, t_1)$, $X(t)$ is regular for $t \in [t_0, t_1)$, $X(t_1)$ is singular such that $X(t_1)c = 0$ for some $c \in \mathbb{C}^n \setminus \{0\}$, and $g(t) := c^T X^T(t) \{Q(t) - \tilde{Q}(t)\} X(t)c = c^T X^T(t) \{U(t) - \tilde{Q}(t)X(t)\}c \equiv 0$ on $[t_0, t_1]$ (since $g(t_0) \geq 0$, $g'(t) \geq 0$ on $[t_0, t_1]$, and $g(t_1) = 0$). Hence $(Q(t) - \tilde{Q}(t))X(t)c \equiv 0$ on $[t_0, t_1]$ (use that $Q(t) \geq \tilde{Q}(t)$), which implies that $\{U(t) - \tilde{Q}(t)X(t)\}c \equiv 0$. Thus $U(t_1)c = X(t_1)c = 0$, contradicting the fact that $\text{rank} \begin{pmatrix} U(t) \\ X(t) \end{pmatrix} \equiv n$. \square

Finally, we state the transformation behaviour of Riccati equations (see [24], p. 104).

PROPOSITION 4. *Given any regular matrix T ; then $Q(t)$ solves (6) on the interval \mathcal{J} if and only if $\tilde{Q}(t) := T^T Q(t) T$ is on \mathcal{J} a solution of $\tilde{Q} + \tilde{A}^T \tilde{Q} + \tilde{Q} \tilde{A} + \tilde{Q} \tilde{B} \tilde{Q} - \tilde{C} = 0$, $\tilde{Q}(t_0) = \tilde{Q}_0$, where $\tilde{A} = T^{-1} A T$, $\tilde{B} = T^{-1} B (T^{-1})^T$, $\tilde{C} = T^T C T$, and $\tilde{Q}_0 = T^T Q_0 T$.*

5. The main result. Before stating our main result, we need two notions, the well-known notion of “controllability” (see, e.g., [5, p. 80] and [2, p. 390]) and a new notion of “strong observability,” which is stronger than the well-known notion of observability [2], [5].

DEFINITION 3. Let be given $(n \times n)$ -matrices A , B , and C . Then we call (i) the pair (A, B) (completely) *controllable*, if $\text{rank} [B, AB, \dots, A^{n-1}B] = n$; and we call (ii) the triple (A, B, C) *strongly observable* if $\dot{x} = Ax + Bu$, $Cx \equiv 0$ on some nondegenerate interval \mathcal{J} , and $u \in C^s(\mathcal{J})$ always implies that $x \equiv 0$ on \mathcal{J} .

Remarks. (i) Obviously, strong observability of (A, B, C) implies that the pair (A, C) is (completely) observable (put $u \equiv 0$). This notion of strong observability seems to be new, and it is actually in a sense equivalent to the asymptotic behaviour of a corresponding Riccati equation (stated in our main result below). This “equivalence” is shown in the next section (Proposition 6). Observe, moreover, that both notions of Definition 3 are *invariant* under transformations according to Proposition 4, i.e., when replacing A, B, C by $\tilde{A} = T^{-1} A T$, $\tilde{B} = T^{-1} B (T^{-1})^T$, and $\tilde{C} = T^T A T$, where T is any regular matrix.

(ii) The assumption of *piecewise continuity* on u (i.e., $u \in C^s(\mathcal{J})$) in Definition 3 may be replaced by the assumption that u is an *entire function* (i.e., $u(t) = \sum_{k=0}^{\infty} u_k t^k$ for all $t \in \mathbb{R}$) without changing the notion of strong observability; we only need this weaker assumption.

Now we can state the main result of this paper.

THEOREM 3. *Let be given (real) $(n \times n)$ -matrices A, B, C, C_0 satisfying the assumptions: B, C , and C_0 are symmetric; B and C_0 are nonnegative definite; the pair (A, B) is controllable; and the triple (A, B, C_0) is strongly observable. Then, for any $t_0 > 0$ and for any symmetric matrix Q_0 , there exists $\lambda_0 = \lambda_0(t_0, Q_0; A, B, C, C_0)$ such that the solution $Q(t; \lambda)$ of the initial value problem*

$$(10) \quad \dot{Q} + A^T Q + Q A + Q B Q - C + \lambda C_0 = 0, \quad Q(0) = Q_0$$

exists on $[0, t_0]$ whenever $\lambda \leq \lambda_0$, and it satisfies

$$(11) \quad \lim_{\lambda \rightarrow -\infty} Q(t_0; \lambda) = \infty; \quad \text{i.e., all eigenvalues of the symmetric } (n \times n)\text{-matrix } Q(t_0; \lambda) \text{ tend to } \infty \text{ as } \lambda \rightarrow -\infty.$$

Proof. Using Corollary 1 and Proposition 3, we may assume that the system is transformed to “normal form.” More precisely, we may suppose that the matrices A ,

B , C , C_0 , and Q_0 have the following form (compare §§ 2 and 3):

$$A = J = \begin{pmatrix} 0 & J_1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & J_{r-1} \\ 0 & \dots & & 0 \end{pmatrix}$$

with $(l_\mu \times l_{\mu+1})$ -matrices $J_\mu = (I \mid 0)$ for $\mu = 1, \dots, r-1$, $B = \mathcal{B} = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$ with the $(l_r \times l_r)$ -identity matrix I , where $1 \leq l_1 \leq \dots \leq l_r$ with $\sum_{\nu=1}^r l_\nu = n$, $C = -\alpha I$, and $Q_0 = -\beta I$ with sufficiently large positive constants α and β . Moreover, we use the notation of § 3: among others, the quantities $\Delta_\nu = l_\nu - l_{\nu-1}$, $m = \sum_{\nu=1}^r \Delta_\nu = l_r$, and the notion of admissible functions $x(t)$ according to Definition 2. Then the assumption of strong observability implies that

$$\int_0^{t_0} x^T(t) C_0 x(t) dt > 0 \quad \text{for all } t_1 > 0 \text{ and for all admissible functions } x(t)$$

(i.e., $\dot{x} = Ax + Bu$ for some function u) on $[0, t_1]$ with $x \not\equiv 0$. Hence Theorem 2 applies, where we may assume that the transformation matrix $T = I$ (using Corollary 2 and again Propositions 3 and 4 such that we can still assume that $A = J$, $B = \mathcal{B}$, $C = -\alpha I$, and $Q_0 = -\beta I$). Now Theorem 2 (noting additionally remark (i) to Theorem 2 and the remark to Corollary 3, below) yields the following: There exists $\delta \geq 1$, and, for a given $0 < \eta \leq t_0$, there exists $\varepsilon_0(\eta) > 0$ such that

$$\begin{aligned} (12) \quad & \int_0^{t_1} x^T(t) C_0 x(t) dt \\ & \geq \int_0^{t_1} \left\{ \sum_{\nu=1}^r \sum_{\mu=l_{\nu-1}+1}^{l_\nu} [\varepsilon_\mu |y_\mu|^2 - \delta \varepsilon_{\mu+1} (|y'_\mu|^2 + \dots + |y_\mu^{(r-\nu)}|)] \right\} dt \\ & = \int_0^{t_1} x^T(t) \tilde{C}_0 x(t) dt \quad \text{with the diagonal matrix} \\ & \tilde{C}_0 = \text{diag} \left(\underbrace{\Delta_1=l_1}_{\varepsilon_1, \dots, \varepsilon_{l_1}}, \underbrace{\Delta_1+\Delta_2=l_2}_{-\delta \varepsilon_2, \dots, -\delta \varepsilon_{l_1+1}, \varepsilon_{l_1+1}, \dots, \varepsilon_{l_2}}, \right. \\ & \quad \left. -\delta \varepsilon_2, \dots, -\delta \varepsilon_{l_2+1}, \varepsilon_{l_2+1}, \dots, \varepsilon_{l_3}, \dots, \right. \\ & \quad \left. \underbrace{-\delta \varepsilon_2, \dots, -\delta \varepsilon_{l_{r-1}+1}, \varepsilon_{l_{r-1}+1}, \dots, \varepsilon_{l_r}}_{\Delta_1+\dots+\Delta_r=l_r} \right) \end{aligned}$$

for all $\eta \leq t_1 \leq t_0$, $0 < \varepsilon_\nu \leq \varepsilon_0$, $\delta \varepsilon_{\nu+1} \leq \varepsilon_\nu$ ($\nu = 1, \dots, m = l_r$ with $\varepsilon_{m+1} := 0$), and for all admissible functions $x(t)$ on $[0, t_0]$. Using this inequality and notation, we obtain the following lemma, which is similar to Proposition 3.

LEMMA. Let $0 < \eta \leq t_0$ such that the solution $Q(t; 0)$ of (10) (with $\lambda = 0$) exists on $[0, \eta]$, and let $\lambda \leq 0$ be fixed. Moreover, assume that $\delta \geq 1$, $\varepsilon_0(\eta) > 0$, $0 < \varepsilon_\nu \leq \varepsilon_0$, $\delta \varepsilon_{\nu+1} \leq \varepsilon_\nu$ for $\nu = 1, \dots, m$, as above. Then, if the solution $\tilde{Q}(t) = \tilde{Q}(t; \lambda)$ of

$$(13) \quad \dot{\tilde{Q}} + A^T \tilde{Q} + \tilde{Q} A + \tilde{Q} B \tilde{Q} - C + \lambda \tilde{C}_0 = 0, \quad \tilde{Q}(0) = Q_0$$

exists on $[0, t_0]$, the solution $Q(t; \lambda)$ of (10) exists on $[0, t_0]$ also, and it satisfies

$$(14) \quad Q(t; \lambda) \geq \tilde{Q}(t; \lambda) \quad \text{for all } t \in [\eta, t_0].$$

Proof of the lemma. First, $Q(t; \lambda)$ exists on $[0, \eta]$ since $\lambda \leq 0$ and $C_0 \geq 0$ by Proposition 3 (with $\tilde{A} = A$, $\tilde{B} = B$, $\tilde{Q}(t) = Q(t; 0)$, $\tilde{C} = C$, and $C - \lambda C_0$ (which is $\geq C$))

instead of C). Then the existence of $Q(t; \lambda)$ on $[\eta, t_0]$ follows similarly as in the second part of the proof of Proposition 3, provided that we have shown (14) (assuming the existence). Now, by Proposition 2, we have that $Q(t) := Q(t; \lambda) = U(t)X^{-1}(t)$, where $X(t)$, $U(t)$ solve the linear system $\dot{X} = AX + BU$, $\dot{U} = (C - \lambda C_0)X - A^T U$ with $X(0) = I$, $U(0) = -\beta I$. The asserted inequality (14) follows, if we show that

$$c^T X^T(t) \{Q(t) - \tilde{Q}(t)\} X(t) c \geq 0 \quad \text{for all } t \in [\eta, t_0] \text{ and } c \in \mathbb{C}^n.$$

Therefore let $c \in \mathbb{C}^n$ and consider that $x(t) = X(t)c$. It follows that $\dot{x} = Ax + Bu$ with $u(t) = U(t)c = Q(t)X(t)c$, i.e., $x(t)$ is admissible, and

$$\frac{d}{dt} x^T(t) \{Q(t) - \tilde{Q}(t)\} x(t) = x^T(t) \{-\lambda(C_0 - \tilde{C}_0)\} x(t) + (\tilde{Q}(t)x(t) - u(t))^T;$$

$$B(\tilde{Q}(t)x(t) - u(t)) \geq x^T(t) \{-\lambda(C_0 - \tilde{C}_0)\} x(t)$$

(using $B \geq 0$, $\dot{x} = Ax + Bu$, and the Riccati equations satisfied by Q , respectively, \tilde{Q}). Hence inequality (12) implies that

$$x^T(t) \{Q(t) - \tilde{Q}(t)\} x(t) \geq (-\lambda) \int_0^t x^T(\tau) (C_0 - \tilde{C}_0) x(\tau) d\tau \geq 0$$

for $t \in [\eta, t_0]$, which yields (14). \square

Now we can continue the proof of Theorem 3. According to the preceding proof and the lemma, we may choose constants $\alpha > 0$, $\beta > 0$, $\eta \in (0, t_0]$, $\varepsilon_0 = \varepsilon_0(\eta) > 0$, and $\delta \geq 1$, and we can put $A = J$, $B = \mathcal{B}$, $C = -\alpha I$, and $Q_0 = -\beta I$. Then it suffices to prove that the solution $\tilde{Q}(t; \lambda)$ of (13) exists on $[0, t_0]$ for $\lambda \leq \lambda_0 < 0$ ($-\lambda_0$ sufficiently large) with $\lim_{\lambda \rightarrow -\infty} \tilde{Q}(t_0; \lambda) = \infty$, where the numbers ε_ν can be chosen as suitable functions of λ , satisfying $0 < \varepsilon_\nu(\lambda) \leq \varepsilon_0$, $\delta \varepsilon_{\nu+1}(\lambda) \leq \varepsilon_\nu(\lambda)$ for $\nu = 1, \dots, m = l_r$. We choose the $\varepsilon_\nu(\lambda)$ inductively as follows: $\varepsilon_1(\lambda) \equiv \varepsilon_0 > 0$ and $\rho_\nu(\lambda) = \sqrt[2r]{-\lambda \varepsilon_\nu}$, $-\lambda \varepsilon_{\nu+1}(\lambda) = \rho_\nu(\lambda)$ for $\nu = 1, \dots, m$. Then

$$(15) \quad \rho_\nu(\lambda) \rightarrow \infty, \quad -\lambda \varepsilon_{\nu+1}(\lambda) \rho_\nu^{-2} = 1/\rho_\nu(\lambda) \rightarrow 0 \quad \text{as } \lambda \rightarrow -\infty \text{ for } \nu = 1, \dots, m$$

such that the restrictions on the ε_ν are satisfied for $\lambda \leq \lambda_0 < 0$. Now we have, by Proposition 2 that $\tilde{Q}(t; \lambda) = U(t)X^{-1}(t)$, where the columns $x(t)$, respectively, $u(t)$, of the $(n \times n)$ -matrices $X(t)$, respectively, $U(t)$, solve the linear system

$$(16) \quad \dot{x} = Jx + \mathcal{B}u, \quad \dot{u} = (-\alpha I - \lambda \tilde{C}_0)x - J^T u.$$

This normal form of (7) is constructed in such a way that it separates into self-adjoint scalar differential equations (which correspond to Sturm-Liouville eigenvalue problems, where the asserted asymptotic behaviour (11) is known; see, e.g., [15, Thm. 11] and [14, Thm. 2]) as follows: Columns $x(t)$, $u(t)$ solve the differential system (15) if and only if $x(t)$ is admissible according to Definition 2 (i.e., $x^T(t) = (x_1^T(t), \dot{x}_1^T(t), \dots, x_r^T(t))$, and we put $\tilde{x}^T(t) = (x_1^T(t), \dots, x_r^T(t)) = (y_1(t), \dots, y_m(t)) \in \mathbb{R}^m$ as in Definition 2), if $u(t)$ is of the form (very abbreviated)

$$u^T(t) = \left((-1)^r y_1^{(2r-1)} + \sum_{\nu=2}^{r-1} (-1)^\nu (-\alpha + \lambda \cdot \delta \varepsilon_2) y_1^{(2\nu-2)}, \dots, y_1(t), \dots, y_m(t) \right),$$

and if the functions $y_1(t), \dots, y_m(t)$ solve the selfadjoint differential equations

$$(17) \quad (-1)^s y_\mu^{(2s)} + \sum_{\nu=1}^{s-1} (-1)^\nu (-\alpha + \lambda \delta \varepsilon_{\mu+1}) y_\mu^{(2\nu)} = (\lambda \varepsilon_\mu + \alpha) y_\mu$$

for $\mu \in \{1, \dots, m\}$, $s \in \{1, \dots, r\}$ with $l_{r-s} < \mu \leq l_{r+1-s}$.

Since the initial value matrix $Q_0 = -\beta I$ is diagonal, the matrices $X(t)$, $U(t)$ "separate completely," which means that $\tilde{Q} = UX^{-1}$ is given by $\tilde{Q}(t; \lambda) = P^T \text{diag}(Q_1(t; \lambda), \dots, Q_m(t; \lambda))P$ with some permutation matrix P , where the diagonal blocks $Q_\mu(t; \lambda)$ solve Riccati equations corresponding to (16). More precisely,

$$(18) \quad \dot{Q}_\mu + A_s^T Q_\mu + Q_\mu A_s + Q_\mu B_s Q_\mu - C_s + \lambda C_{0s} = 0, \quad Q_\mu(0) = -\beta I$$

for $\mu \in \{1, \dots, m\}$, $s \in \{1, \dots, r\}$, $l_{r-s} < \mu \leq l_{r+1-s}$, where the $(s \times s)$ -matrices A_s , B_s , C_s , C_{0s} are given by $C_s = -\alpha I$, $C_{0s} = \text{diag}(\varepsilon_\mu, -\delta\varepsilon_{\mu+1}, \dots, -\delta\varepsilon_{\mu+1})$, $B_s = \text{diag}(0, \dots, 0, 1)$, and the companion matrix

$$A_s = \begin{pmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}.$$

These solutions $Q_\mu(t; \lambda)$ are evaluated asymptotically in [15, Thm. 11], [14, Thm. 2], and the corresponding result needed here is stated in Proposition 5, below. Since $-\lambda\varepsilon_\mu = \rho_\mu^{2r}(\lambda) \geq \rho_\mu^{2s}(\lambda) \rightarrow \infty$ and $-\lambda\varepsilon_{\mu+1} = \rho_\mu$, Proposition 5 (with $\varepsilon(\rho_\mu) = 1/\rho_\mu$) implies that $Q_\mu(t; \lambda)$ exists on $[0, t_0]$ for $\lambda \leq \lambda_0$ and that $\lim_{\lambda \rightarrow \infty} Q_\mu(t_0; \lambda) = \infty$ for all $\mu = 1, \dots, m$, which yields the desired assertion. \square

PROPOSITION 5. *Assume that A is the $(s \times s)$ -companion matrix, that $B = \text{diag}(0, \dots, 0, 1)$ as above, and that $C_0 = \text{diag}(1, 0, \dots, 0)$, $C = \text{diag}(r_0(\rho), \dots, r_{s-1}(\rho))$ such that*

$$|r_\nu(\rho)|\rho^{2(\nu-s)} \leq \varepsilon(\rho) \rightarrow 0 \quad \text{as } \rho \rightarrow \infty \text{ for } \nu = 0, \dots, s-1.$$

Then, for any $t_0 > 0$ and for any symmetric matrix Q_0 , the solution $Q(t; \rho)$ of

$$\dot{Q} + A^T Q + Q A + Q B Q + C + \rho^{2s} C_0 = 0, \quad Q(0) = Q_0$$

exists on $[0, t_0]$ for $\rho > 0$ sufficiently large, and it satisfies

$$Q(t_0; \rho) = \rho \tilde{D}_\rho \{G + O(\varepsilon(\rho))\} \tilde{D}_\rho \rightarrow \infty \text{ as } \rho \rightarrow \infty,$$

where $\tilde{D}_\rho = \text{diag}(\rho^{s-1}, \dots, \rho, 1)$ and where G is a positive definite matrix (depending on s only).

Proof. The proof (and the assertion) of Theorem 2 in [14] uses the more restrictive assumption that $\varepsilon(\rho) = O(\rho^{-2})$ (which is needed there for Theorem 1). An inspection of the explicit calculations in that proof, however, immediately shows that Proposition 5, above (with $O(\varepsilon(\rho))$ instead of $O(\rho^{-2})$ in [14, Thm. 2]), is true. \square

6. Applications to the linear regulator in optimal control. In this section, we prove a consequence of Theorem 3 concerning the quotient of quadratic functionals

$$R(x) := \int_0^{t_0} \{x^T C x + u^T B u\} dt \Big/ \int_0^{t_0} x^T C_0 x dt,$$

where the denominator is positive for admissible $x(t) \not\equiv 0$. Then the quotient $R(x)$ is $\geq \lambda_0$ if and only if the quadratic functional

$$I(x; \lambda_0) := \int_0^{t_0} \{x^T C x + u^T B u\} dt - \lambda_0 \int_0^{t_0} x^T C_0 x dt$$

is ≥ 0 , which relates the minimization of $R(x)$ to the optimal linear regulator problem. Our result on the minimization of $R(x)$ is well known if $B = C_0 = I$ (see, e.g., [23, Chap. IV, Thm. 3.1] and [3, § 7.2]); in particular, the method of proof is normally used when dealing with the linear regulator in optimal control (see, e.g., [17, § 3.3]).

Moreover, Theorem 3 can be used to prove a far more general result (including general boundary conditions and also time-dependent systems) via Theorem 4 of [3]. This will be done in a forthcoming paper, where a general oscillation result (compare [3, Thm. 3]) will be derived. Here, however, we need a rather special case, which was already used in the inductive proof of the central inequality (2) in Theorem 2, above. Moreover, the proof below is rather easy (when Theorem 3 is known), and it makes this paper self-contained.

THEOREM 4. *Under the assumptions of Theorem 3, there exists*

$$\lambda_0 := \min \{R(x) \mid x(t) \text{ admissible and } \neq 0\},$$

where $x(t)$ is admissible if $\dot{x} = Ax + Bu$ for some $u \in C^s[0, t_0]$, and where

$$R(x) := \int_0^{t_0} \{x^T Cx + u^T Bu\} dt \Big/ \int_0^{t_0} x^T C_0 x dt.$$

Moreover, $\lambda_0 = R(z)$, where

$$\dot{z} = Az + Bv, \quad \dot{v} = (C - \lambda_0 C_0)z - A^T v \quad \text{on } [0, t_0], \quad v(0) = v(t_0) = 0.$$

Proof. Consider the solution $Q(t; \lambda)$ of the initial value problem (10) with $Q_0 = 0$. Then, by Proposition 2, we have that $Q(t; \lambda) = U(t; \lambda)X^{-1}(t; \lambda)$ where $\dot{X} = AX + BU$, $\dot{U} = (C - \lambda C_0)X - A^T U$ and $X(0) = I$, $U(0) = 0$. By Theorem 3 and [3, Thm. 1 with $S_{24} = I$, $S_{13} = 0$], there exists $\lambda_0 > -\infty$ such that the following holds: $Q(t; \lambda)$ exists (i.e., $X(t; \lambda)$ is regular) on $[0, t_0]$ for $\lambda \leq \lambda_0$, $Q(t_0; \lambda)$ is positive definite for $\lambda < \lambda_0$, and $Q(t_0; \lambda_0)$ is singular, i.e., $U(t_0; \lambda_0)c = 0$ for some $c \in \mathbb{R}^n \setminus \{0\}$ (observe that $\lambda_0 < \infty$ is an immediate consequence of inequality (19), below). Now assume that $x(t)$ is admissible, i.e., $\dot{x} = Ax + Bu$ with $u \in C^s[0, t_0]$. Then we obtain from Riccati's equation (10) that

$$(19) \quad \frac{d}{dt} x^T(t) Q(t; \lambda) x(t) = -(u - Qx)^T B(u - Qx) + u^T Bu + x^T (C - \lambda C_0) x$$

for $t \in [0, t_0]$ and $\lambda \leq \lambda_0$ (observe that formula (18) reduces to the so-called "Picone identity" [3, (6.1)] in this special case). Since B is nonnegative definite, $Q(0) = 0$, and, since $Q(t_0; \lambda)$ is positive definite for $\lambda < \lambda_0$, we obtain the inequality

$$(20) \quad \int_0^{t_0} \{x^T Cx + u^T Bu\} dt \geq \lambda \int_0^{t_0} x^T C_0 x dt \quad \text{for } \lambda < \lambda_0.$$

Hence $R(x) \geq \lambda_0$. Moreover, $\lambda_0 = R(z)$, where $z(t) = X(t; \lambda_0)c \neq 0$ satisfies $\dot{z} = Az + Bv$, $\dot{v} = (C - \lambda_0 C_0)z - A^T v$ for $v(t) = U(t; \lambda_0)c$ with $v(0) = v(t_0) = 0$, which completes the proof. \square

Remarks. (i) Observe that controllability and strong observability imply that the differential system $\dot{x} = Ax + Bu$, $\dot{u} = (C - \lambda C_0)x - A^T u$ is *normal* according to [3, (A2)], which is needed for the application of [3, Thm. 1], above. Moreover, if $C_0 = I$, then this notion of normality coincides with "identically normal" in [23, p. 313] and with condition $[C]$ in [8, pp. 36, 37].

(ii) Next, note that $C_0 \geq 0$ and strong observability is equivalent to $\int_0^{t_0} x^T C_0 x dt > 0$ for every admissible and not vanishing $x(t)$ and every $t_0 > 0$. Hence the denominator of the quotient $R(x)$ is never zero when taking the minimum as above.

(iii) A simple but nontrivial application of Theorem 4 (which is not covered by known results, to our knowledge) results from the following example.

Example. Let

$$C_0 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and let C be any symmetric (4×4) -matrix, e.g., $C = -\alpha I$. Then (A, B) is controllable, and $x(t) = (x_i(t)) \in C_1^s[0, t_0]$ is admissible if and only if $\dot{x}_1 = x_3$ and $\dot{x}_2 = x_4$. Hence, $\int_0^{t_0} x^T C_0 x \, dt = \int_0^{t_0} \{x_2^2 + (x_1 + \dot{x}_2)^2\} \, dt > 0$ for all admissible $x(t) \neq 0$. This implies that (A, B, C_0) is strongly observable. For this example, the full strength of Theorem 2 (in particular, Case 2.1 of its proof) is needed for the proof of Theorem 3, which becomes, in this rather elementary case, by no means trivial (including Theorem 4).

An immediate consequence of Theorem 4 is the following result.

COROLLARY 3. *Under the assumptions of Theorem 3, suppose additionally that $\int_0^t \{x^T C x + u^T B u\} \, d\tau > 0$ for all $t > 0$ and all admissible $x(\tau)$ with $x(\tau) \neq 0$ (and $\dot{x} = Ax + Bu$). Then there exists (a maximal) $\alpha = \alpha(t) > 0$ such that*

$$\int_0^t \{x^T C x + u^T B u\} \, d\tau \geq \alpha \int_0^t x^T C_0 x \, d\tau$$

for all admissible $x(\tau) \in C_1^s[0, t]$.

Remark. Of course, $\alpha(t)$, above, depends on t in general; a rather elementary reasoning either directly with $R(x)$ or considering $\det U(t; \lambda)$ (according to the proof of Theorem 4) shows that $\alpha(t)$ is continuous on $(0, \infty)$. Hence, for $0 < \delta \leq t_0$, there exists $\varepsilon = \varepsilon(\delta, t_0) > 0$ such that $\alpha(t) \geq \varepsilon$ for all $t \in [\delta, t_0]$.

Finally, we prove that strong observability is necessary for the statement of Theorem 3 (and actually also of Theorem 4!).

PROPOSITION 6. *Let there be given (real) $(n \times n)$ -matrices A, B, C, C_0 , and Q_0 such that B, C, C_0 , and Q_0 are symmetric, and such that B is nonnegative definite. Assume that, for any $t_0 > 0$, the solution $Q(t; \lambda)$ of (10) exists on $[0, t_0]$ whenever $\lambda \leq \lambda_0(t_0)$ and satisfies (11). Then C_0 is nonnegative definite, and (A, B, C_0) is strongly observable.*

Proof. Let there be given $t_0 > 0$ and an admissible function $x(t)$ on $[0, t_0]$ with $x(t_0) \neq 0$ such that $\dot{x} = Ax + Bu$ for some $u \in C^s[0, t_0]$. Since B is nonnegative definite and since the solution $Q(t; \lambda)$ of (10) exists on $[0, t_0]$ for $\lambda \leq \lambda_0$, we obtain from (18) the inequality

$$\begin{aligned} (-\lambda) \int_0^{t_0} x^T C_0 x \, dt &\geq -x^T(0) Q_0 x(0) - \int_0^{t_0} \{u^T B u + x^T C x\} \, dt \\ &\quad + x^T(t_0) Q(t_0; \lambda) x(t_0) \quad \text{for all } \lambda \leq \lambda_0. \end{aligned}$$

Now $x^T(t_0) Q(t_0; \lambda) x(t_0) \rightarrow \infty$ as $\lambda \rightarrow -\infty$ by our assumption, and this implies that $\int_0^{t_0} x^T C_0 x \, dt > 0$. Hence C_0 is nonnegative definite, and (A, B, C_0) is strongly observable (compare remark (ii) after Theorem 4). \square

REFERENCES

- [1] B. D. O. ANDERSON AND D. G. LUENBERGER, *Design of multivariable feedback systems*, Proc. IEE, 114 (1967), pp. 395-399.
- [2] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

- [3] G. BAUR AND W. KRATZ, *A general oscillation theorem for selfadjoint differential systems with applications to Sturm–Liouville eigenvalue problems and quadratic functionals*, Rend. Circ. Mat. Palermo (2), 38 (1989), pp. 329–370.
- [4] F. BRAUER, *Singular selfadjoint boundary value problems for the differential equation $Lx = \lambda Mx$* , Trans. Amer. Math. Soc., 88 (1958), pp. 331–345.
- [5] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [6] C. T. CHEN, *A note on pole assignment*, IEEE Trans. Automat. Control, 13 (1968), pp. 597–598.
- [7] C. K. CHUI AND G. CHEN, *Linear Systems and Optimal Control*, Springer, Berlin, New York, 1989.
- [8] W. A. COPPEL, *Disconjugacy*, Lecture Notes in Math., Springer, Berlin, New York, 1971.
- [9] E. J. DAVISON, *On pole assignment in multivariable linear systems*, IEEE Trans. Automat. Control, 13 (1968), pp. 747–748.
- [10] O. HIJAB, *Stabilization of Control Systems*, Springer, Berlin, New York, 1987.
- [11] E. KAMKE, *Über die definiten selbstadjungierten Eigenwertaufgaben bei gewöhnlichen linearen Differentialgleichungen I–IV*, Math. Z. 45 (1939), pp. 759–787; 46 (1940), pp. 231–250, 251–286; 48 (1942), pp. 67–100.
- [12] H. W. KNOBLOCH AND H. KWAKERNAAK, *Lineare Kontrolltheorie*, Springer, Berlin, New York, 1980.
- [13] W. KRATZ, *A limit theorem for matrix-solutions of Hamiltonian systems*, Rend. Circ. Mat. Palermo (2), 36 (1987), pp. 457–473.
- [14] ———, *Asymptotic behaviour of Riccati's differential equation associated with self-adjoint scalar equations of even order*, Czech. Math. J., 38 (113) (1988), pp. 351–365.
- [15] W. KRATZ AND A. PEYERIMHOFF, *A treatment of Sturm–Liouville eigenvalue problems via Picone's identity*, Analysis, 5 (1985), pp. 97–152.
- [16] K. KREITH, *Oscillation Theory*, Lecture Notes in Math., Springer, Berlin, New York, 1970.
- [17] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, John Wiley, New York, 1972.
- [18] H. LEPTIN, *Kamkesche Eigenwertprobleme und Hilbert–Schmidt-Operatoren*, Math. Z., 82 (1963), pp. 133–151.
- [19] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 290–293.
- [20] F. J. MÄRKER, *Zur Asymptotik gewisser Riccatischer Matrix-Differentialgleichungen*, Ph.D. dissertation, University of Ulm, Ulm, Germany, 1990.
- [21] M. MORSE, *The Calculus of Variations in the Large*, AMS Colloquium Publications, Vol. 18, Providence, RI, 1934.
- [22] ———, *Variational analysis: Critical extremals and Sturmian extensions*, John Wiley, New York, 1973.
- [23] W. T. REID, *Ordinary Differential Equations*, John Wiley, New York, 1971.
- [24] ———, *Riccati Differential Equations*, Academic Press, New York, 1972.
- [25] F. W. SCHÄFKE AND A. SCHNEIDER, *S-hermitesche Rand- und Eigenwert-probleme I/II*, Math. Ann., 162 (1965), pp. 9–26; 165 (1966), pp. 236–260.
- [26] R. E. SKELTON, *Dynamic Systems Control*, John Wiley, New York, 1988.
- [27] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 660–665.

NONLINEAR OPTIMIZATION: ON CONNECTED COMPONENTS OF LEVEL SETS*

HUBERTUS TH. JONGEN† AND JAN RUECKMANN‡

Abstract. This paper is concerned with differentiable constrained optimization problems in finite dimensions. The constraint qualification used is of Mangasarian–Fromovitz type. This paper studies the connected component of the level set of the objective function containing a specific local minimizer, considering that component as a function of the objective level. Special attention is paid to compactness and continuity aspects, also in connection with the occurrence of stationary points. Moreover, a covering of such a compact connected component is presented with differentiable curves tending to the specific local minimizer along which the objective function monotonically decreases.

Key words. level sets, connected components, descent curves, Mangasarian–Fromovitz constraint qualification, stationary points, strongly stable stationary points, MFCQ

AMS(MOS) subject classification. 90C31

1. Introduction. Main results. Let R^n be the n -dimensional Euclidean space, and $C^k(R^n, R)$, $k \geq 0$, the space of real-valued, k -times continuously differentiable functions on R^n . If $f \in C^1(R^n, R)$, then $Df(x)$ represents the derivative (row vector) of f at x . We consider the following optimization problem in standard form:

$$(P) \quad \text{Minimize } f(x) \text{ subject to } M,$$

where the feasible set M is defined as

$$M = \{x \in R^n \mid h_i(x) = 0, i \in I, g_j(x) \leq 0, j \in J\}$$

and $I = \{1, \dots, m\}$, $J = \{1, \dots, s\}$ are two finite index sets with $m < n$. Furthermore, we assume that $f, h_i, g_j \in C^2(R^n, R)$, $i \in I$, and $j \in J$. For $x \in R^n$, we define the set of active inequality constraints

$$J_0(x) = \{j \in J \mid g_j(x) = 0\}.$$

We say that a point $\hat{x} \in M$ is a *stationary point* of (P) if there exist numbers u_i , $i \in I \cup J_0(\hat{x})$ satisfying the following system:

$$(1.1) \quad Df(\hat{x}) + \sum_{i \in I} u_i Dh_i(\hat{x}) + \sum_{j \in J_0(\hat{x})} u_j Dg_j(\hat{x}) = 0, \quad u_j \geq 0, j \in J_0(\hat{x}).$$

By $E(P)$, we denote the set of all stationary points of (P). If \bar{x} is a local minimizer and if, in addition, a certain constraint qualification is fulfilled at \bar{x} , then $\bar{x} \in E(P)$ (cf. [3]). In this study, we need two constraint qualifications, the well-known linear independence constraint qualification (LICQ) and the Mangasarian–Fromovitz constraint qualification (MFCQ) at a point \hat{x} . These are defined below.

* Received by the editors July 25, 1990; accepted for publication (in revised form) August 21, 1991.

† Aachen University of Technology, Department of Mathematics, Templergraben 55, W-5100 Aachen, Germany.

‡ Leipzig University of Technology, Department of Mathematics and Computer Science, P.O. Box 66, O-7030 Leipzig, Germany.

LICQ is said to hold at \hat{x} if the vectors $Dh_i(\hat{x}), i \in I, Dg_j(\hat{x}), j \in J_0(\hat{x})$ are linearly independent.

MFCQ is said to hold at \hat{x} if the following two conditions are satisfied:

- (i) The vectors $Dh_i(\hat{x}), i \in I$, are linearly independent; and
- (ii) There exists a vector $\xi \in R^n$ satisfying

$$Dh_i(\hat{x})\xi = 0, \quad i \in I; \quad Dg_j(\hat{x})\xi < 0, \quad j \in J_0(\hat{x}).$$

It is easily seen that LICQ implies MFCQ at \hat{x} . In recent years, many results belonging to stability or continuity analysis have been proved under the assumption that MFCQ is satisfied at certain points. Recall, for example, that MFCQ is satisfied at $\hat{x} \in E(P)$ if and only if the set of Lagrange multipliers, $u_i, i \in I \cup J_0(\hat{x})$ satisfying (1.1) is bounded [4]. Furthermore, we refer to Kojima's necessary and sufficient conditions for strong stability of stationary points under MFCQ [10]. In [6], [7], [11], one-parametric optimization problems are studied, and, in particular, the local structure of the one-dimensional set of stationary points. There it is shown that the failure of MFCQ gives rise to bifurcation points of the feasible sets under consideration and to boundary points of the set of stationary points. Under MFCQ, the (structural) stability and continuity of the set-valued mapping

$$(h_i, g_j, i \in I, j \in J) \mapsto \{x \in R^n \mid h_i(x) = 0, i \in I, g_j(x) \leq 0, j \in J\}$$

are studied in papers [5], [9]. Finally, we also refer to continuity aspects given in [13].

Many of these results are basic for the theoretical background and the development of concepts and algorithms in nonlinear optimization. In this paper, we study properties of a connected component of the level set of f depending on the level. Special attention is paid to compactness and continuity aspects, also in connection with the occurrence of stationary points. We call a subset $A \subset B$ a *connected component* of B if A is connected and if, for every connected subset C of B with $A \subset C$, it holds that $A = C$. A set A is called *path-connected* if, for any two points $x, y \in A$, there exists a continuous mapping $\omega: t \in [0, 1] \mapsto \omega(t) \in A$ with $\omega(0) = x$ and $\omega(1) = y$. In particular, the set $\{\omega(t) \mid t \in [0, 1]\}$ is called a *path*.

For $a \in R$, we define $M^a = \{x \in M \mid f(x) \leq a\}$, $M_0^a = \{x \in M \mid f(x) < a\}$ and, for $\bar{x} \in M$, $a \geq f(\bar{x})$ ($a > f(\bar{x})$), the set $M(\bar{x}, a)$ ($M_0(\bar{x}, a)$) as the connected component of M^a (M_0^a), which contains \bar{x} . To compute a stationary point of (P), we can use an algorithm that starts with a point x^0 and computes a sequence $\{x^i\}$ whose cluster points belong to $E(P)$. In many cases, this sequence $\{x^i\}_{i=0,1,2,\dots}$ belongs to (a neighbourhood of) exactly one connected component of M (or one path-connected component of M , if using trajectory methods), and the condition $f(x^{i+1}) \leq f(x^i) - \varepsilon^i$ ($\varepsilon^i > 0, \varepsilon^i \rightarrow 0$) is satisfied. Therefore we are interested in (topological) properties of the sets $M(\bar{x}, a)$ depending on a . We investigate properties of $M(\bar{x}, a)$, $a \in [f(\bar{x}), \bar{a}]$, and, throughout the paper, we assume the following two conditions:

- (V1) \bar{x} is an isolated local minimizer of (P) and \bar{a} is fixed with $\bar{a} > f(\bar{x})$; and
- (V2) MFCQ is satisfied at all points $x \in M(\bar{x}, \bar{a})$.

Before stating the first theorem, we return to Kojima's strong stability. In [10] the so-called stationary index (s.i. (\hat{x})) of a strongly stable stationary point \hat{x} , the appropriate generalization of the Morse index [12], is defined if MFCQ is satisfied at \hat{x} . In particular, if MFCQ holds at \hat{x} , but LICQ does not, then s.i. (\hat{x}) = 0.

The close relationship between the change of the topological structure of M^a when passing the level of a strongly stable stationary point \hat{x} , and the stationary index s.i. (\hat{x}) is investigated in [8] and [5, Thm. C]. A consequence of MFCQ and strong stability is the following theorem.

THEOREM 1. *Suppose that every point of $(M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}) \cap E(P)$ is strongly stable. Then $M(\bar{x}, \bar{a})$ is path-connected.*

The next theorem provides the existence of a covering of the set $M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}$ with disjoint trajectories contained in $M(\bar{x}, \bar{a})$ and tending to \bar{x} along which f monotonically decreases.

THEOREM 2. *Let $M(\bar{x}, \bar{a})$ be compact and $M(\bar{x}, \bar{a}) \cap E(P) = \{\bar{x}\}$. Then for each $\hat{x} \in M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}$, there exists a continuous function*

$$\Phi_{\hat{x}}: t \in [0, f(\hat{x}) - f(\bar{x})] \mapsto \Phi_{\hat{x}}(t) \in M(\bar{x}, \bar{a})$$

with the following properties:

- (i) $\Phi_{\hat{x}}|_{(0, f(\hat{x}) - f(\bar{x}))}$ is continuously differentiable;
- (ii) $\Phi_{\hat{x}}(0) = \hat{x}$, $\Phi_{\hat{x}}(f(\hat{x}) - f(\bar{x})) = \bar{x}$;
- (iii) $f(\Phi_{\hat{x}}(t)) = f(\hat{x}) - t$;
- (iv) Let $\tilde{x} \in M(\bar{x}, \bar{a})$ with $f(\tilde{x}) \geq f(\hat{x})$. Then either (a) for each $t \in [0, f(\hat{x}) - f(\tilde{x})]$, $\Phi_{\hat{x}}(t) = \Phi_{\tilde{x}}(f(\tilde{x}) - f(\hat{x}) + t)$, or (b) for each $t \in [0, f(\hat{x}) - f(\bar{x})]$, $\Phi_{\hat{x}}(t) \neq \Phi_{\tilde{x}}(f(\tilde{x}) - f(\hat{x}) + t)$.

For our purposes, we introduce the following assumption.

- (V3) The stationary points in $M(\bar{x}, \bar{a})$, other than \bar{x} itself, constitute a finite set, each element of which is strongly stable with stationary index strictly greater than 1 (in the sense of Kojima). If this set is nonempty, we denote it by $\{x^1, \dots, x^p\}$.

For our purposes, we define the cone $N_{\hat{x}}$ at $\hat{x} \in M$ as follows:

$$N_{\hat{x}} = \left\{ \sum_{i \in I} \lambda_i Dh_i(\hat{x}) - \sum_{j \in J_0(\hat{x})} \mu_j Dg_j(\hat{x}) \left| \begin{array}{l} \lambda_i \in \mathbb{R}, i \in I \\ \mu_j \in \mathbb{R}, j \in J_0(\hat{x}) \\ \mu_j \geq 0 \end{array} \right. \right\} \cup \{0\}.$$

Furthermore, let $\|\cdot\|$ denote the Euclidean norm and, for $\varepsilon > 0$, $A \subset \mathbb{R}^n$ and $\hat{x} \in \mathbb{R}^n$, put

$$d(\hat{x}, A) = \inf \{\|\hat{x} - x\| \mid x \in A\},$$

$$B_\varepsilon(\hat{x}) = \{x \in \mathbb{R}^n \mid \|x - \hat{x}\| < \varepsilon\} \quad \text{and}$$

$$B_\varepsilon(A) = \{x \in \mathbb{R}^n \mid d(x, A) < \varepsilon\}.$$

The condition defined below is an appropriate modification of the ("Palais-Smale") condition C^* in [5].

DEFINITION 1. Condition C^+ is fulfilled by the function $g \in C^1(\mathbb{R}^n, \mathbb{R})$ on a set $S \subset \mathbb{R}^n$ if

$$\inf \{d(Dg(x), N_x) \mid x \in S\} > 0.$$

If (V3) is fulfilled, then there exist open bounded neighbourhoods U_0 of \bar{x} and U_i of x^i , $i = 1, \dots, p$ such that

$$U_i \cap U_j = \emptyset, \quad i \neq j, \quad i, j = 0, \dots, p \quad \text{and}$$

$$(1.2) \quad \text{MFCQ is satisfied at all } x \in \bigcup_{i=0}^p U_i.$$

Henceforth, we sometimes use, under (V3), the following additional assumption:

- (V4) The function f fulfils condition C^+ on the set $M(\bar{x}, \bar{a}) \setminus \bigcup_{i=0}^p U_i$ for every choice of open bounded neighbourhoods U_0 of \bar{x} and U_i of x^i , $i = 1, \dots, p$ satisfying (1.2).

THEOREM 3. *Let conditions (V3) and (V4) be fulfilled. Then, for every $a \in [f(\bar{x}), \bar{a}]$,*

$$(1.3) \quad M(\bar{x}, \bar{a}) \cap M^a = M(\bar{x}, a).$$

In the following two theorems, we consider continuity properties of the point-to-set mapping

$$\Gamma: a \in [f(\bar{x}), \bar{a}] \mapsto \Gamma(a) = M(\bar{x}, a).$$

A well-detailed study of continuity aspects related to point-to-set mappings is given in [1]. Using the definitions from [1], the mapping Γ is said to be *continuous at $\hat{a} \in [f(\bar{x}), \bar{a}]$* if Γ is (i) lower semicontinuous in the sense of Berge (lsc-B) at \hat{a} , as well as (ii) upper semicontinuous in the sense of Hausdorff (usc-H) at \hat{a} . Γ is called lsc-B at \hat{a} if, for any open set $\Omega \subset \mathbb{R}^n$ with $\Omega \cap \Gamma(\hat{a}) \neq \emptyset$, there exists a neighbourhood $B_\varepsilon(\hat{a})$ such that $\Omega \cap \Gamma(a) \neq \emptyset$ for each $a \in B_\varepsilon(\hat{a})$. We say that Γ is usc-H at \hat{a} if, for each $\varepsilon > 0$, there exists a $\delta > 0$ such that $\Gamma(a) \subset B_\varepsilon(\Gamma(\hat{a}))$ for every $a \in B_\delta(\hat{a})$.

Before stating the next theorem, we remark that Theorem 4 is true without assuming (V2).

THEOREM 4. (i) *If Γ is lsc-B at \bar{a} , then*

$$(1.4) \quad \text{cl } M_0(\bar{x}, \bar{a}) = M(\bar{x}, \bar{a})$$

(where cl denotes the closure);

(ii) *Let $M_0(\bar{x}, \bar{a})$ be path-connected and let (1.4) be satisfied. Then Γ is lsc-B at \bar{a} ;*

(iii) *If $M(\bar{x}, \bar{a})$ is compact, then Γ is usc-H at \bar{a} .*

From the latter theorem, we obtain the following results for our special situation.

THEOREM 5. *Assume conditions (V3) and (V4). Then*

(i) $\text{cl } M_0(\bar{x}, a) = M(\bar{x}, a)$ for all $a \in (f(\bar{x}), \bar{a}]$;

(ii) $M(\bar{x}, a)$ is compact for all $a \in [f(\bar{x}), \bar{a}]$;

(iii) Γ is continuous at a for all $a \in [f(\bar{x}), \bar{a}]$.

In § 2 we present some basic lemmas used later. Then § 3 contains the proof of the Theorems 1–5 given above, as well as some additional remarks.

2. Lemmas, preliminary results. In [5, Thm. A] it is shown that M is a topological manifold with the boundary $\delta M = \{x \in \mathbb{R}^n \mid h_i(x) = 0, i \in I, \max_{j \in J} g_j(x) = 0\}$ if MFCQ is satisfied at all points of M . As a local consequence, we formulate the following lemma.

Lemma 1. *Let MFCQ be satisfied at a point $\tilde{x} \in M$. Then*

(a) *There exist open neighbourhoods $U(\tilde{x})$ of \tilde{x} and \tilde{U} of the origin in \mathbb{R}^{n-m} such that $U(\tilde{x}) \cap M$ is connected and homeomorphic to \tilde{U} if $J_0(\tilde{x}) = \emptyset$, and otherwise homeomorphic to $\tilde{U} \cap \mathbb{R}^{n-m-1} \times \mathbb{R}_+$, where $\mathbb{R}_+ = \{y \in \mathbb{R} \mid y \geq 0\}$;*

(b) *Let $U(\tilde{x})$ be chosen as in (a). Then $U(\tilde{x}) \cap M$ is path-connected.*

Proof. If $J_0(\tilde{x}) = \emptyset$, then MFCQ (i) implies that, locally around \tilde{x} , the set M is a C^2 -manifold and therefore homeomorphic to a (path-connected) neighbourhood \tilde{U} of the origin in \mathbb{R}^{n-m} . If $J_0(\tilde{x}) \neq \emptyset$, the desired homeomorphism is constructed in the proof of [5, Thm. A]. \square

Corollary 1. *Let A be a connected component of M , and let MFCQ be satisfied at all points of A . Then A is path-connected.*

Proof. The proof is immediate from the fact that A is connected and locally path-connected.

In the following lemma, let $\eta: \mathcal{O} \rightarrow \mathbb{R}^n$ be a vector field defined on the open subset $\mathcal{O} \subset \mathbb{R}^n$. The lemma is a well-known result from the theory of ordinary differential equations.

LEMMA 2. (a) If η is locally Lipschitzian, respectively, belongs to the class C^k ($k \geq 1$), then, for each $\tilde{x} \in \mathcal{O}$, there exist an open neighbourhood $U(\tilde{x})$ of \tilde{x} , a number $\varepsilon(\tilde{x}) > 0$, and a unique mapping $\Phi \in C^k(U(\tilde{x}) \times (-\varepsilon(\tilde{x}), \varepsilon(\tilde{x})), \mathcal{O})$ ($k = 0$, respectively, $k \geq 1$) such that

$$\frac{\partial \Phi}{\partial t}(x, t) = \eta(\Phi(x, t)), \quad \eta(x, 0) = x, \quad \text{and}$$

$$\Phi_{t_1+t_2}(x) = \Phi_{t_1} \circ \Phi_{t_2}(x)$$

for all $x \in U(\tilde{x})$, $\{t_1, t_2\} \subset (-\varepsilon(\tilde{x}), \varepsilon(\tilde{x}))$ satisfying $|t_1 + t_2| < \varepsilon(\tilde{x})$, where we put $\Phi_t(\cdot) = \Phi(\cdot, t)$.

(b) If $\mathcal{O} = R^n$, η locally Lipschitzian (or C^k ($k \geq 1$)), and bounded, then Φ is defined on $R^n \times R$; i.e., η is completely integrable, and $\Phi(\cdot, t)$ is a homeomorphism (or C^k -diffeomorphism) for each $t \in R$.

In the following lemma, we need condition C^+ introduced in § 1.

Lemma 3. Let \hat{M} be a closed subset of M and let MFCQ be satisfied at all points of \hat{M} . Suppose that f fulfils condition C^+ on \hat{M} . Then there exists a bounded vector field $\eta \in C^1(R^n, R^n)$ with the following properties:

(i) For all $x \in M$,

$$Df(x)\eta(x) \leq 0,$$

$$Dh_i(x)\eta(x) = 0, \quad i \in I,$$

$$Dg_j(x)\eta(x) \leq 0, \quad j \in J_0(x);$$

(ii) For all $x \in \hat{M}$,

$$Df(x)\eta(x) \leq -1;$$

(iii) The flow Φ of the vector field η satisfies the relation

$$\Phi(\cdot, t)[M(\bar{x}, a)] \subset M(\bar{x}, a)$$

for all $a \geq f(\bar{x})$ and all $t \geq 0$.

Proof. Lemma 3 can be proved in the same way as [5, Thm. C], substituting \hat{M} for the set M_a^b in the latter reference. In particular, condition C^+ yields the boundedness of the vector field η .

In the next lemma, let X be a topological Hausdorff space and denote by $H_q(X)$, $q = 0, 1, 2, \dots$ the q th singular homology space over the reals with respect to X . Roughly speaking, the dimension of $H_q(X)$ represents the number of “ $(q + 1)$ -dimensional holes” in X (see [8, Chap. 5]). The next lemma is well known (cf. [15]).

LEMMA 4. (a) The cardinality of path-connected components of X is equal to $\dim H_0(X)$ (where \dim denotes the dimension);

(b) Let X and Y be homotopy-equivalent Hausdorff spaces. Then, for all $q = 0, 1, 2, \dots$, $H_q(X) \cong H_q(Y)$.

3. Proofs of the theorems and remarks.

Proof of Theorem 1. The proof is given in two steps.

Step 1. First, we show that every nonempty connected component Z of $M(\bar{x}, \bar{a}) \setminus \{x \in E(P) | f(x) = \bar{a}\}$ is path-connected. Let $\tilde{x} \in Z$ and

$$A^0 = \{x \in Z | \text{there exists a path in } Z \text{ connecting } x \text{ and } \tilde{x}\}.$$

We know $\tilde{x} \in A^0$, and we prove that A^0 is open and closed in Z . Suppose that $\hat{x} \in A^0$.

Since $\hat{x} \notin \{x \in E(P) | f(x) = \bar{a}\}$, MFCQ is fulfilled at \hat{x} with respect to the set

$$M^{\bar{a}} = \left\{ x \in R^n \left| \begin{array}{l} h_i(x) = 0, i \in I \\ g_j(x) \leq 0, j \in J \\ f(x) \leq \bar{a} \end{array} \right. \right\}.$$

Now Lemma 1 implies the existence of an open neighbourhood $U(\hat{x})$ of \hat{x} such that $U(\hat{x}) \cap \{x \in E(P) | f(x) = \bar{a}\} = \emptyset$ and $U(\hat{x}) \cap M^{\bar{a}}$ is path-connected. The latter two facts imply that $U(\hat{x}) \cap M^{\bar{a}} \subset Z$ and $U(\hat{x}) \cap M^{\bar{a}} = U(\hat{x}) \cap Z$, as well as $U(\hat{x}) \cap Z \subset A^0$. Now let $\bar{x}^i \in A^0$, $\bar{x}^i \rightarrow x'$ and $x' \in Z$. From Lemma 1 and the fact that $x' \notin \{x \in E(P) | f(x) = \bar{a}\}$, we see that $\bar{x}^{i_0} \in U(x')$ for an appropriate index i_0 , and, consequently, there exists a path in Z connecting x' and \bar{x}^{i_0} .

Step 2. Now let $\tilde{x} \in M(\bar{x}, \bar{a}) \cap E(P)$ and $f(\tilde{x}) = \bar{a}$. From (V1), we know that $f(\bar{x}) < \bar{a}$, and hence $\tilde{x} \neq \bar{x}$. If the stationary index of \tilde{x} vanishes, then \tilde{x} is a strict local minimizer. It follows that there exists a neighbourhood U of \tilde{x} such that $M^{\bar{a}} \cap U = \{\tilde{x}\}$. Consequently, the connectedness of $M(\bar{x}, \bar{a})$ implies that $\tilde{x} \notin M(\bar{x}, \bar{a})$. So the stationary index of \tilde{x} must be greater than zero, and hence LICQ is satisfied at \tilde{x} (cf. [10, Thm. 7.2]). Consequently, there exists a vector $v \in R^n$ satisfying

$$Dh_i(\tilde{x})v = 0, \quad i \in I, \quad Dg_j(\tilde{x})v = 0, \quad j \in J_0(\tilde{x}), \quad v^T D^2 L(\tilde{x})v < 0.$$

Here $L = f + \sum_{i \in I} u_i h_i + \sum_{j \in J_0(\tilde{x})} u_j g_j$ is the corresponding Lagrange function (with unique numbers u_i, u_j such that (1.1) at \tilde{x} is satisfied). Furthermore, choose vectors $\xi_l \in R^n, l = 1, \dots, n - |I| - |J_0(\tilde{x})|$ such that $Dh_i(\tilde{x})^T, i \in I, Dg_j(\tilde{x})^T, j \in J_0(\tilde{x}), \xi_l, l = 1, \dots, n - |I| - |J_0(\tilde{x})|$ form a basis of R^n . So we can construct the local C^2 -coordinate transformation

$$\Omega(x) = \begin{pmatrix} h_i(x), i \in I \\ g_j(x), j \in J_0(\tilde{x}) \\ \xi_l^T(x - \tilde{x}), l = 1, \dots, n - |I| - |J_0(\tilde{x})| \end{pmatrix}$$

and the C^1 -vector field $v(x) = D\Omega^{-1}(\Omega(x))D\Omega(\tilde{x})v$, defined on an appropriate neighbourhood of \tilde{x} . It is easily seen that there exists a neighbourhood $\tilde{U}(0)$ of 0 such that the relations

$$\Phi(\tilde{x}, t) \in M \setminus E(P) \quad \text{and} \quad f(\Phi(\tilde{x}, t)) < \bar{a}$$

hold for all $t \in \tilde{U}(0) \setminus \{0\}$, where $\Phi(\tilde{x}, t)$ denotes the trajectory through \tilde{x} corresponding to the vector field $v(x)$.

Now Theorem Cb. in [5] implies that locally around \tilde{x} the set $M(\bar{x}, \bar{a}) \setminus \{\tilde{x}\}$ is connected if s.i. $(\tilde{x}) \geq 1$. In this case, the path $\{\Phi(\tilde{x}, t) | t \geq 0, t \in \tilde{U}(0)\}$ connects the points $\Phi(\tilde{x}, 0) = \tilde{x}$ and $\Phi(\tilde{x}, t) \in M(\bar{x}, \bar{a}) \setminus E(P), t > 0, t \in \tilde{U}(0)$. If s.i. $(\tilde{x}) = 1$, then, locally around \tilde{x} , the set $M(\bar{x}, \bar{a}) \setminus \{\tilde{x}\}$ consists of two connected components; a moment of reflection shows that the sets

$$\{\Phi(\tilde{x}, t) | t > 0, t \in \tilde{U}(0)\} \quad \text{and} \quad \{\Phi(\tilde{x}, t) | t < 0, t \in \tilde{U}(0)\}$$

do not belong to the same connected component.

The results given in Steps 1 and 2 imply that the set

$$\{x \in M(\bar{x}, \bar{a}) | \text{there exists a path in } M(\bar{x}, \bar{a}) \text{ connecting } \bar{x} \text{ and } x\}$$

is nonempty, open, and closed in $M(\bar{x}, \bar{a})$. Thus we are done. \square

Proof of Theorem 2. Let $\hat{x} \in M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}$ be arbitrarily chosen and fixed. As a consequence of $\hat{x} \notin E(P)$, condition (V2), and the well-known Farkas lemma, we see

that the set

$$(3.1) \quad A(\hat{x}) = \left\{ v \in R^n \mid \begin{array}{l} Df(\hat{x})v < 0, Dh_i(\hat{x})v = 0, i \in I \\ Dg_j(\hat{x})v < 0, j \in J_0(\hat{x}) \end{array} \right\}$$

is nonempty. Now choose $\xi_l \in R^n, l = m+1, \dots, n$ in such a way that the vectors $Dh_i(\hat{x}), i \in I, \xi_l, l = m+1, \dots, n$ form a basis of R^n . Put $y = \Omega_{\hat{x}}(x)$, where $y_i = h_i(x), i \in I$ and $y_l = \xi_l(x - \hat{x}), l = m+1, \dots, n$. Obviously, or $v(\hat{x}) \in A(\hat{x})$, there exists an open neighbourhood $\mathcal{O}(\hat{x})$ of \hat{x} such that, for any $x \in \mathcal{O}(\hat{x})$, the Jacobian $D\Omega_{\hat{x}}(x)$ is regular and $J_0(x) \subset J_0(\hat{x})$, as well as $\hat{v}(x) \in A(x)$, where $\hat{v}(x) = D\Omega_{\hat{x}}^{-1}(\Omega_{\hat{x}}(x))D\Omega_{\hat{x}}(\hat{x})v(\hat{x})$ and $A(x)$ as in (3.1). Now we can define the open covering

$$\mathcal{O} = \bigcup_{\bar{x} \in M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}} \mathcal{O}(\hat{x})$$

of the set $M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}$ with $\bar{x} \notin \mathcal{O}$.

By means of a C^1 -partition of unity subordinate to the covering $\{\mathcal{O}(\hat{x}), \hat{x} \in M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}\}$, selecting on each $\mathcal{O}(\hat{x})$ the vector field $\hat{v}(x)$, we get a C^1 -vector field $\tilde{F}(x)$ on \mathcal{O} satisfying $\tilde{F}(x) \in A(x)$ for every $x \in M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}$. In particular, $Df(x)\tilde{F}(x) \neq 0, x \in \mathcal{O}$ and, therefore, we can define the C^1 -vector field

$$F(x) = \frac{\tilde{F}(x)}{\|Df(x)\tilde{F}(x)\|}$$

on \mathcal{O} satisfying $F(x) \in A(x)$ and $Df(x)F(x) = -1$ for every $x \in M(\bar{x}, \bar{a}) \setminus \{\bar{x}\}$. In virtue of Lemma 2, there exist an $\varepsilon(\hat{x}) > 0$ and an open neighbourhood $V(\hat{x}) \times (-\varepsilon(\hat{x}), \varepsilon(\hat{x}))$ of $(\hat{x}, 0)$, as well as a unique flow $\Phi(x, t) \in C^1$ of $F(x)$ on $V(\hat{x}) \times (-\varepsilon(\hat{x}), \varepsilon(\hat{x}))$ with $\Phi(x, 0) = x$ for $x \in V(\hat{x})$. Obviously, we have that

$$(3.2) \quad f(\Phi(x, t_1)) - f(\Phi(x, t_2)) = t_2 - t_1$$

for any $\{t_1, t_2\} \subset (-\varepsilon(\hat{x}), \varepsilon(\hat{x}))$ with $|t_1 + t_2| < \varepsilon(\bar{x})$ and $x \in V(\hat{x})$.

PROPOSITION. $\Phi(\hat{x}, \cdot)$ is defined for all $t \in [0, f(\hat{x}) - a]$ and every $a \in (f(\bar{x}), f(\hat{x}))$.

We proceed with the proof, assuming that the above proposition has already been proved. By $\{\Phi(\hat{x}, t) \mid t \in [0, f(\hat{x}) - f(\bar{x})]\} \subset M(\bar{x}, \bar{a})$ and the compactness of $M(\bar{x}, \bar{a})$, it follows that the set

$$\Phi(\hat{x}, f(\hat{x}) - f(\bar{x})) := \left\{ x \in M(\bar{x}, \bar{a}) \mid \begin{array}{l} \text{there exists a sequence } \{t^i\} \subset [0, f(\hat{x}) - f(\bar{x})] \\ \text{satisfying } \lim_{t^i \rightarrow f(\hat{x}) - f(\bar{x})} \Phi(\hat{x}, t^i) = x \end{array} \right\}$$

is nonempty and that it is $f(x) = f(\bar{x})$ for every $x \in [\Phi(\hat{x}, f(\hat{x}) - f(\bar{x}))]$. In view of $M(\bar{x}, \bar{a}) \cap M^{f(\bar{x})} = \{\bar{x}\}$, we obtain $\lim_{t \rightarrow f(\hat{x}) - f(\bar{x})} \Phi(\hat{x}, t) = \bar{x}$.

Finally, we must prove the above proposition.

Proof. Let $a \in (f(\bar{x}), f(\hat{x}))$ be arbitrarily chosen and fixed. Since the set $M(\bar{x}, a, \bar{a}) := \{x \in M(\bar{x}, \bar{a}) \mid f(x) \geq a\}$ is compact, there exists a finite covering $\{\tilde{V}(\bar{x}^i) \mid \bar{x}^i \in M(\bar{x}, a, \bar{a}), i \in \tilde{I}\}$ of $M(\bar{x}, a, \bar{a})$, where \tilde{I} is a finite index set and $\tilde{V}(\bar{x}^i)$ (as above) are open neighbourhoods of $\bar{x}^i, i \in \tilde{I}$. Put $\varepsilon = \min \{\varepsilon(\bar{x}^i) \mid i \in \tilde{I}\}$. Since $\Phi(x, t)$ is uniquely determined and well defined on $M(\bar{x}, a, \bar{a}) \times (-\varepsilon, \varepsilon)$, it can be uniquely extended at every reached point from $M(\bar{x}, a, \bar{a})$.

By (3.2) we obtain that

$$f(\Phi(\hat{x}, t)) = f(\hat{x}) - t, \quad t \in [0, f(x) - a],$$

and we achieve the desired result. \square

Remark 1. If we delete the compactness of $M(\bar{x}, \bar{a})$ in Theorem 2 without any substitute, then the trajectory $\Phi(\hat{x}, t), t \geq 0$, constructed in the proof of Theorem 2, need not tend to \bar{x} , and it might happen that $\lim_{t \rightarrow \bar{t}} \|\Phi(\hat{x}, t)\| = \infty$ for some \bar{t} .

In Fig. 1, such a situation is sketched for the case where $|I| = |J| = \emptyset$. Here the function f is represented by means of some level lines, and the dotted, boundary line is identified with the "point at infinity."

Remark 2. In the following, we occasionally use the expression: "We add locally at x^i a constant ε to the function f " ($i \in \{1, \dots, p\}$ and fixed). By this expression, we mean that we add to f a function $\varepsilon \cdot \xi \in C^2(R^n, R)$ having the following properties:

- (i) $0 \leq \xi(x) \leq 1$,
- (ii) $\xi(x) = 1$ in a neighbourhood of x^i ,
- (iii) There exists an open bounded neighbourhood \bar{U}^i of x^i with support $(\xi) \subset \bar{U}^i \cap U^i$, where U^i is defined as in (1.2), and x^i is the only stationary point of the resulting perturbed problem in \bar{U}^i and is strongly stable with stationary index s.i. (x^i) .

Proof of Theorem 3. Obviously, $M(\bar{x}, \bar{a}) \cap M^a = M(\bar{x}, \bar{a})$. Now suppose there exists an $a \in (f(\bar{x}), \bar{a})$ such that (1.3) is not satisfied. We add locally at every $x^i, i = 1, \dots, p$ with $f(x^i) = \bar{a}$ a constant $-\varepsilon_1 < 0$ and at every $x^i, i = 1, \dots, p$ with $f(x^i) = a$ a constant $\varepsilon_2 > 0$ to f . We denote this perturbed function by \tilde{f} . Furthermore, put $\tilde{M}^a = \{x \in M \mid \tilde{f}(x) \leq a\}$, and $\tilde{M}(\bar{x}, a)$ is the connected component of \tilde{M}^a containing \bar{x} .

It is easily seen that we can choose the perturbation functions and the constants ε_1 and ε_2 in such a way that the following properties are fulfilled:

- (a) $a + \varepsilon_2 < \bar{a} - \varepsilon_1$,
- (b) $\tilde{M}(\bar{x}, a) \subset M(\bar{x}, a)$, $M(\bar{x}, \bar{a}) \subset \tilde{M}(\bar{x}, \bar{a})$,
- (c) There exists a point $\hat{x} \in \{M(\bar{x}, \bar{a}) \cap M^a\} \setminus M(\bar{x}, a)$ with $\hat{x} \in \tilde{M}^a$,
- (d) $\tilde{M}(\bar{x}, \bar{a}) \subset M(\bar{x}, \bar{a}) \cup \bigcup_j \bar{U}^j$, where \bar{U}^j is defined as in Remark 2 and j varies in the set $\{j \in \{1, \dots, p\} \mid f(x^j) = \bar{a}\}$. By (V3) and $\tilde{f}(x^i) < \bar{a}, i = 1, \dots, p$, it follows that MFCQ is satisfied at all $x \in \tilde{M}(\bar{x}, \bar{a})$ with respect to the constraints $h_i(x) = 0, i \in I, g_j(x) \leq 0, j \in J$, and $\tilde{f}(x) \leq \bar{a}$.

Properties (a)–(c) imply that $\hat{x} \in (\tilde{M}(\bar{x}, \bar{a}) \cap \tilde{M}^a) \setminus \tilde{M}(\bar{x}, a)$, and hence

$$(3.3) \quad \tilde{M}(\bar{x}, \bar{a}) \cap \tilde{M}^a \neq \tilde{M}(\bar{x}, a).$$

By property (d) and Lemma 1, we see that $\tilde{M}(\bar{x}, \bar{a})$ is a topological manifold, and, consequently, there exists an open neighbourhood V of $\tilde{M}(\bar{x}, \bar{a})$ satisfying $V \cap \tilde{M}^a = \tilde{M}(\bar{x}, \bar{a})$. Then $\{V, R^n \setminus \tilde{M}(\bar{x}, \bar{a})\}$ is an open covering of R^n . Let $\Phi_i \in C^1(R^n, R), i = 1, 2$ be a partition of unity subordinate to this covering. In particular, we have that

$$\Phi_1|_{\tilde{M}(\bar{x}, \bar{a})} \equiv 1 \quad \text{and} \quad \Phi_2|_{R^n \setminus V} \equiv 1.$$

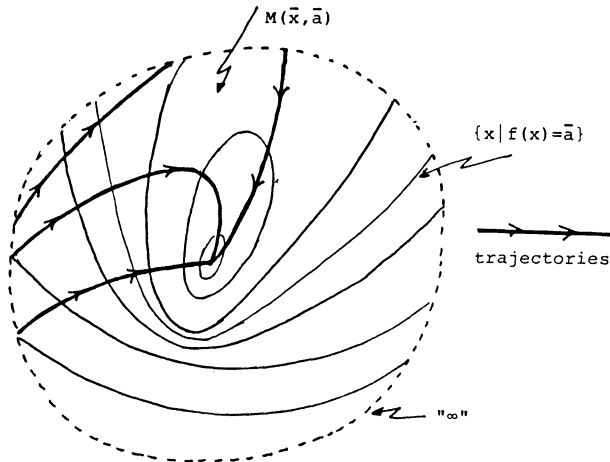


FIG. 1

Now we consider the following optimization problem. Minimize $\tilde{f}(x)$ subject to \hat{M} , where

$$\hat{M} = \left\{ x \in R^n \left| \begin{array}{l} h_i(x) = 0, i \in I \\ \Phi_1(x)g_j(x) + \Phi_2(x) \leq 0, j \in J \end{array} \right. \right\}.$$

Put $\hat{M}^a = \{x \in \hat{M} \mid \tilde{f}(x) \leq a\}$. Obviously, we have that

$$(3.4) \quad \hat{M}^{\bar{a}} = \tilde{M}(\bar{x}, \bar{a}), \quad \hat{M}^a = \tilde{M}(\bar{x}, \bar{a}) \cap \tilde{M}^a,$$

and the stationary points of this problem are \bar{x}, x^1, \dots, x^p . Also, x^i is strongly stable with the stationary index s.i. (x^i) , $i = 1, \dots, p$. Furthermore, by (d), MFCQ is satisfied at all $x \in \hat{M}^{\bar{a}}$, and (V4) is also satisfied (if we substitute f by \tilde{f} and $M(\bar{x}, \bar{a})$ by $\tilde{M}(\bar{x}, \bar{a})$). These facts remain true if we perturb \tilde{f} in such a way that $\tilde{f}(x^i) \neq \tilde{f}(x^j)$ for $i \neq j$, $i, j = 1, \dots, p$, and if \hat{M}^a and $\hat{M}^{\bar{a}}$ are not changed. Then, by Theorem Cb in [5] and Theorem 5.2.1 in [8], it is $\dim H_0(\hat{M}^a) = \dim H_0(\hat{M}^{\bar{a}}) (=1)$, and so, in view of Lemma 4(a), \hat{M}^a and $\hat{M}^{\bar{a}}$ have the same number of path-connected components. This, however, is in contradiction with (3.3), (3.4), and the definition of $\tilde{M}(\bar{x}, a)$. The validity of (1.3) at $a = f(\bar{x})$ is obvious, since we can essentially use a similar argument. \square

Remark 3. In the following two examples, $|I| = |J| = \emptyset$, and f is sketched by means of some level lines. These examples illustrate that (1.3) is, generally, not fulfilled if one of the conditions (V3), (V4) is violated. In Fig. 2, the connected set $M(\bar{x}, \bar{a})$ contains a stationary point with *stationary index* 1, and $M(\bar{x}, \bar{a}) \cap M^a$ consists of two connected components for $a < \bar{a}$ with a sufficiently near \bar{a} .

In Fig. 3, condition (V4) is not fulfilled, and, as in Fig. 1, the dotted boundary line is identified with the "point at infinity."

Proof of Theorem 4. Throughout this proof, the index i always varies in the whole set of natural numbers: $i = 1, 2, 3, \dots$

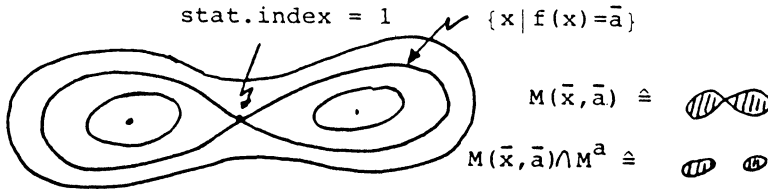


FIG. 2

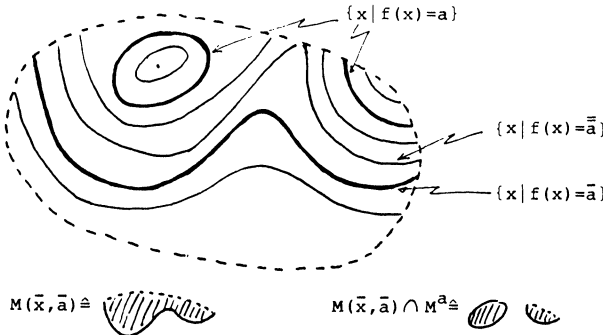


FIG. 3

(i) Obviously, we have that $\text{cl } M_0(\bar{x}, \bar{a}) \subset M(\bar{x}, \bar{a})$, and, therefore, we suppose that there exists an \hat{x} satisfying

$$\hat{x} \in M(\bar{x}, \bar{a}) \setminus \text{cl } M_0(\bar{x}, \bar{a}).$$

Let $\{\varepsilon^i\}$ be a sequence with $\varepsilon^i \rightarrow 0$, $\varepsilon^i > 0$. Since $M(\bar{x}, \bar{a}) \cap B_{\varepsilon^i}(\hat{x}) \neq \emptyset$ and Γ is lsc-B at \bar{a} , there exists numbers $\delta^i > 0$ such that $M(\bar{x}, \bar{a}) \cap B_{\varepsilon^i}(x) \neq \emptyset$ for all $a \in B_{\delta^i}(\bar{a})$. Thus we have sequences $\{\bar{x}^i\}, \{a^i\}$ with $a^i < \bar{a}$, $\bar{x}^i \in M(\bar{x}, a^i) \cap B_{\varepsilon^i}(\hat{x})$, and, consequently, $\bar{x}^i \in M_0(\bar{x}, \bar{a})$ (since $M(\bar{x}, a^i)$ is connected, $\bar{x} \in M(\bar{x}, a^i)$, $M(\bar{x}, a^i) \subset M_0^{\bar{a}}$), $\bar{x}^i \rightarrow \hat{x}$, and $\hat{x} \in \text{cl } M_0(\bar{x}, \bar{a})$, a contradiction.

(ii) Suppose that there exists an open set \mathcal{O} with $M(\bar{x}, \bar{a}) \cap \mathcal{O} \neq \emptyset$ and a sequence $\{a^i\}$ such that $a^i \rightarrow \bar{a}$ and $M(\bar{x}, a^i) \cap \mathcal{O} = \emptyset$. Since $M(\bar{x}, a_1) \subset M(\bar{x}, a_2)$ whenever $a_1 \leq a_2$, it follows that $a^i < \bar{a}$ and

$$(3.5) \quad M(\bar{x}, a) \cap \mathcal{O} = \emptyset \quad \text{for all } a < \bar{a}.$$

By (3.5) and assumption (1.4), there is an $\hat{x} \in M_0(\bar{x}, \bar{a}) \cap \mathcal{O}$ with $f(\hat{x}) = \hat{a} \geq f(\bar{x})$ and $\hat{x} \notin M(\bar{x}, \hat{a})$. Now let $\tilde{x} \in M(\bar{x}, \hat{a})$ (note that $M(\bar{x}, \hat{a}) \neq \emptyset$ since $\bar{x} \in M(\bar{x}, \hat{a})$), and hence $\tilde{x} \in M_0(\bar{x}, \bar{a})$. In view of the path-connectedness of $M_0(\bar{x}, \bar{a})$, there exists a path \mathcal{C} in $M_0(\bar{x}, \bar{a})$ connecting \tilde{x} and \hat{x} , and therefore $\hat{a} \leq \max \{f(x) \mid x \in \mathcal{C}\} < \bar{a}$ and $\hat{x} \in M(\bar{x}, \max \{f(x) \mid x \in \mathcal{C}\}) \cap \mathcal{O}$, which contradicts (3.5).

(iii) Suppose that there exist a positive number $\bar{\varepsilon}$ and sequences $\{a^i\}, \{\bar{x}^i\}$ such that $a^i > \bar{a}$, $a^i \rightarrow \bar{a}$, $\bar{x}^i \in M(\bar{x}, a^i)$ and $\bar{x}^i \notin B_{\bar{\varepsilon}}(M(\bar{x}, \bar{a}))$. Furthermore, let $\varepsilon^i > 0$ and $\varepsilon^i \rightarrow 0$. Since $B_{\varepsilon^i}(M(\bar{x}, a^i))$ is open and connected and, therefore, path-connected, we have, for every i , a path \mathcal{C}^i in $B_{\varepsilon^i}(M(\bar{x}, a^i))$ connecting \bar{x}^i and \bar{x} . By $\tilde{\mathcal{C}}^i$ we denote the connected component of $\mathcal{C}^i \cap \text{cl } B_{\bar{\varepsilon}}(M(\bar{x}, \bar{a}))$, which contains \bar{x} . As a consequence of $\bar{x}^i \notin B_{\bar{\varepsilon}}(M(\bar{x}, \bar{a}))$, there is an $\hat{x}^i \in \tilde{\mathcal{C}}^i$ such that $d(\hat{x}^i, M(\bar{x}, \bar{a})) = \bar{\varepsilon}$, and the compactness of $M(\bar{x}, \bar{a})$ implies the existence of a cluster point \hat{x} of $\{\hat{x}^i\}$. We show that the set

$$A = \{y \mid \text{there exists a subsequence } \{y^{\nu_i}\}_{\{\nu_i\} \subset \{i\}} \text{ with } y^{\nu_i} \in \tilde{\mathcal{C}}^{\nu_i} \text{ and } y^{\nu_i} \rightarrow y\}$$

is connected. Then, by $\tilde{\mathcal{C}}^i \subset B_{\varepsilon^i}(M(\bar{x}, a^i))$ and $\bar{x} \in A$, we obtain that $d(y^{\nu_i}, M(\bar{x}, a^{\nu_i})) \rightarrow 0$, $f(y) \leq \bar{a}$ for $y \in A$, and, consequently, $A \subset M(\bar{x}, \bar{a})$. This, however, cannot be true in view of $\hat{x} \in A$ and $d(\hat{x}, M(\bar{x}, \bar{a})) = \bar{\varepsilon}$. If A were not connected, then there would exist open sets $\mathcal{O}_1, \mathcal{O}_2$ satisfying $\mathcal{O}_1 \cap \mathcal{O}_2 = \emptyset$, $\mathcal{O}_1 \cap A \neq \emptyset$, $\mathcal{O}_2 \cap A \neq \emptyset$ and $A \subset \mathcal{O}_1 \cup \mathcal{O}_2$. Now let $\bar{x} \in \mathcal{O}_1$ and $\tilde{x} \in A \cap \mathcal{O}_2$. Then there exists a sequence $\{\tilde{x}^{\nu_i}\}$ such that $\tilde{x}^{\nu_i} \rightarrow \tilde{x}$ and $\tilde{x}^{\nu_i} \in \tilde{\mathcal{C}}^{\nu_i} \cap \mathcal{O}_2$. By $\bar{x} \in \tilde{\mathcal{C}}^{\nu_i} \cap \mathcal{O}_1$ and since $\tilde{\mathcal{C}}^{\nu_i}$ is connected, it follows that $\tilde{\mathcal{C}}^{\nu_i} \not\subset \mathcal{O}_1 \cup \mathcal{O}_2$, and thus there is an $\hat{x}^{\nu_i} \in \tilde{\mathcal{C}}^{\nu_i} \setminus (\mathcal{O}_1 \cup \mathcal{O}_2)$. The compactness of $M(\bar{x}, \bar{a})$ implies that A contains a cluster point \hat{x} of $\{\hat{x}^{\nu_i}\}$ satisfying $\hat{x} \in A \setminus (\mathcal{O}_1 \cup \mathcal{O}_2)$, which contradicts $A \subset \mathcal{O}_1 \cup \mathcal{O}_2$. This completes the proof. \square

Remark 4. The proof of Theorem 4(ii) shows that the condition that $M_0(\bar{x}, \bar{a})$ is path-connected can be weakened. It is sufficient to assume that, for any point $\hat{x} \in M_0(\bar{x}, \bar{a}) \setminus \text{cl } \bigcup_{a < \bar{a}} M(\bar{x}, a)$, there exists a point $\tilde{x} \in \bigcup_{a < \bar{a}} M(\bar{x}, a)$ and a path in $M_0(\bar{x}, \bar{a})$ connecting \hat{x} and \tilde{x} .

The following example, however, illustrates that the path-connectedness of $M_0(\bar{x}, \bar{a})$ cannot be deleted without any substitute. In Fig. 4(a), f is sketched by means of some level lines, where \bar{a} is the maximum of f and the line $[x^1, x^2]$ connecting x^1 and x^2 is just $\{x \mid f(x) = \bar{a}\}$. Outside of the "boundary level lines," let f be constant with $f \equiv f(\tilde{x})$. The feasible set M is described by the inequalities $0 \leq x_1 \leq x_1^0, x_2^0 \leq x_2 \leq$

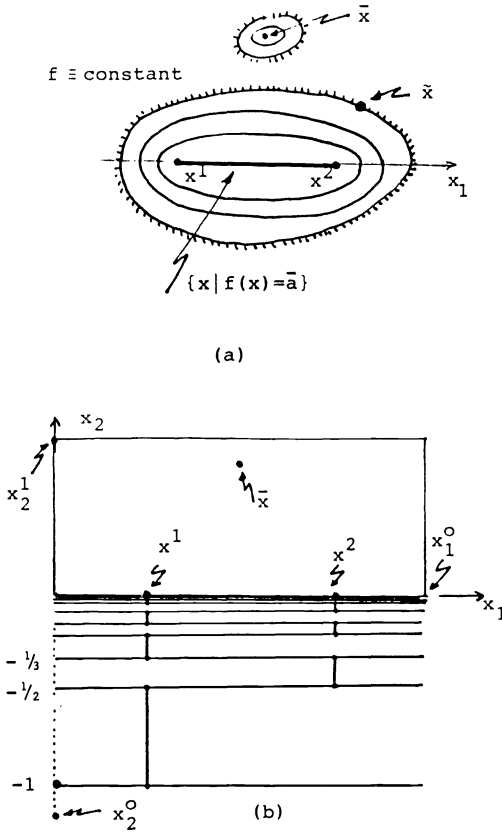


FIG. 4

x_2^1 , and $g(x) \leq 0$, where

$$\begin{aligned} \{x \mid g(x) \leq 0\} &= \{x \in \mathbb{R}^2 \mid x_2 \geq 0\} \cup \bigcup_{n \in \mathbb{N}} \left\{ \left(x_1, -\frac{1}{n} \right) \mid x_1 \in \mathbb{R} \right\} \\ &\quad \cup \bigcup_{k \in \mathbb{N}} \{x^1\} \times \left[-\frac{1}{2k-1}, -\frac{1}{2k} \right] \\ &\quad \cup \bigcup_{k \in \mathbb{N}} \{x^2\} \times \left[-\frac{1}{2k}, -\frac{1}{2k+1} \right] \end{aligned}$$

(cf. Fig. 4(b)). The existence of such a smooth function g follows from the fact that every nonempty closed subset A of \mathbb{R}^n can be represented as the zero set of a nonnegative function $h_A \in C^\infty(\mathbb{R}^n, \mathbb{R})$ (cf. [2]). (Recall that Theorem 4 is true without assuming (V2).)

The sets $M(\bar{x}, \bar{a})$ ($=M$) and $M_0(\bar{x}, \bar{a}) = M(\bar{x}, \bar{a}) \setminus [x^1, x^2]$ are connected, but *not path-connected*. For $a < \bar{a}$, the set $M(\bar{x}, a)$ contains only points with nonnegative x_2 -components and so we have that $\text{cl } M_0(\bar{x}, \bar{a}) = M(\bar{x}, \bar{a})$, but Γ is not lsc-B at \bar{a} .

Remark 5. In Theorem 4(iii), we proved that Γ is usc-H at \bar{a} if $M(\bar{x}, \bar{a})$ is compact. For the function f sketched in Fig. 3, it holds that

$$\text{cl } M_0(\bar{x}, \bar{a}) = M(\bar{x}, \bar{a}).$$

We see that $M(\bar{x}, \bar{a})$ is not compact and that Γ is not usc-H at \bar{a} .

Proof of Theorem 5. (i) Let $a \in (f(\bar{x}), \bar{a}]$ be arbitrarily chosen and fixed. We show the inclusion $M(\bar{x}, a) \subset \text{cl } M_0(\bar{x}, a)$, since $\text{cl } M_0(\bar{x}, a) \supset M(\bar{x}, a)$ is obvious. Let $x \in M(\bar{x}, a)$, and we distinguish two cases. If $x \notin E(P)$, then (V2) and Lemma 1 imply the existence of a sequence $\{\bar{x}^i\}$ (the index i is always varying in the whole set of natural numbers) such that $\bar{x}^i \rightarrow x$ and $\bar{x}^i \in M_0^a \cap M(\bar{x}, a)$. Then, in view of Theorem 3, we obtain that $\bar{x}^i \in M_0(\bar{x}, a)$. Now let $x \in E(P)$. If $x = \bar{x}$, we can put $\bar{x}^i = \bar{x}$, and it is $\bar{x}^i \in M_0(\bar{x}, a)$ and $\bar{x}^i \rightarrow x$. For $x \neq \bar{x}$, by Theorem 1 and (V3), there exists a sequence $\{\hat{x}^i\} \subset M(\bar{x}, a)$ satisfying $\hat{x}^i \rightarrow x$ and $\hat{x}^i \notin E(P)$. Using the argumentation of the first case ($x \notin E(P)$) for each \hat{x}^i , there exists, for every $\varepsilon > 0$ and every index i_0 , a point $\tilde{x}(i_0, \varepsilon) \in M_0(\bar{x}, a)$ such that $\|\tilde{x}(i_0, \varepsilon) - \hat{x}^{i_0}\| < \varepsilon$. Consequently, for a sequence $\{\varepsilon^i\}$ with $\varepsilon^i > 0$, $\varepsilon^i \rightarrow 0$, we can construct a sequence $\{\tilde{x}^i\} \subset M_0(\bar{x}, a)$ such that $\|\tilde{x}^i - x\| < \varepsilon^i$, and therefore $\tilde{x}^i \rightarrow x$.

(ii) The proof of (ii) is given in four steps, which are demonstrated below.

Step 1. In view of Lemma 2.2 in [14], there exists an $a^+ > f(\bar{x})$ such that $M(\bar{x}, a^+) \cap E(P) = \{\bar{x}\}$ and $M(\bar{x}, a)$ is compact for each $a \in [f(\bar{x}), a^+]$. By (V3) we can choose a discretization $f(\bar{x}) = a_0 < a^+ < a_1 < \dots < a_r = \bar{a}$ with $1 \leq r-1 \leq p$ such that, for each $i \in \{1, \dots, p\}$, there exists an index $j \in \{1, \dots, r\}$ satisfying $f(x^i) = a_j$.

Step 2. Choose an open bounded neighbourhood \hat{U}_0 of \bar{x} such that $f(x) < a^+$ for all $x \in \hat{U}_0 \cap M$. For any $\hat{a} \in (a^+, a^1)$, we put $\hat{M} = M(\bar{x}, \hat{a}) \setminus \hat{U}_0$, and, in virtue of (V4), Lemma 3, and Theorem 3, we obtain a C^1 -vector field η on R^n with the flow $\psi(x, t)$ satisfying

$$\psi(\cdot, \hat{a} - a^+)[\hat{M}] \subset M(\bar{x}, a^+).$$

Since $\psi(\cdot, a - a^+)$ is a homeomorphism and, according to Step 1, $M(\bar{x}, a^+)$ is compact, it follows that \hat{M} is compact, and, considering the boundedness of \hat{U}_0 , $M(\bar{x}, a)$ is compact for each $a \in [f(\bar{x}), a^1]$.

Step 3. Obviously, we can choose positive numbers γ, ε and the neighbourhoods $U_i, i = 0, \dots, p$ in (V4) in such a way that $\gamma + 3\varepsilon < \min\{a_k - a_{k-1} | k = 1, \dots, r\}$ and

$$(3.6) \quad f(x) - f(x^i) < \gamma \quad \text{for each } x \in U_i.$$

Without loss of generality, suppose that $\gamma = a^+ - f(\bar{x})$. In this step, let $k = 1, a_k < \bar{a}$. i is always varying in the set $A_k = \{i = 1, \dots, p | f(x_i) = a_k\}$. Now, we locally add at x^i the constant -3ε to f and define \hat{U}_i as an open bounded neighbourhood of x^i satisfying $\tilde{f}(x) < a_k - 2\varepsilon$ whenever $x \in \hat{U}_i$, where \tilde{f} is the so-perturbed function f . Furthermore, $\tilde{M}(\bar{x}, a)$ denotes, for $a \geq f(\bar{x})$, the connected component of $\{x \in M | f(x) \leq a\}$, which contains \bar{x} . Then, for each $a \in [f(\bar{x}), \bar{a}]$, it is $M(\bar{x}, a) \subset \tilde{M}(\bar{x}, a)$, and some consideration shows that, in view of (3.6) and Theorem 3, we also have that $M(\bar{x}, a) \subset M(\bar{x}, a) \cup \bigcup_{i \in A_k} U_i$. For $a \in [f(\bar{x}), \bar{a}]$, it follows that $M(\bar{x}, a)$ is compact if and only if $\tilde{M}(\bar{x}, a)$ is compact, and, in particular, Step 2 simplifies the compactness of $\tilde{M}(\bar{x}, a)$, $a \in [f(\bar{x}), a_1]$. For an arbitrarily chosen $\hat{a} \in [a_k, a_{k+1})$, we put

$$\hat{M} = \tilde{M}(\bar{x}, \hat{a}) \setminus \bigcup_{\{i=1, \dots, p | f(x_i) \leq a_k\} \cup \{0\}} \hat{U}_i,$$

and, by (V4), Lemma 3, and Theorem 3, we obtain a C^1 -vector field $\tilde{\eta}(x)$ with the flow $\tilde{\psi}(x, t)$. The above construction implies that

$$\tilde{\psi}(\cdot, \hat{a} - (a_k - \varepsilon))[\hat{M}] \subset \tilde{M}(\bar{x}, a_k - \varepsilon).$$

Since $\tilde{\psi}(\cdot, \hat{a} - (a_k - \varepsilon))$ is a homeomorphism and $\tilde{M}(\bar{x}, a_k - \varepsilon)$ is compact and

$$\bigcup_{\{i=1, \dots, p | f(x_i) \leq a_k\} \cup \{0\}} \hat{U}_i$$

is bounded, the compactness of the sets $\hat{M}, \tilde{M}(\bar{x}, \hat{a})$, and $M(\bar{x}, a)$ follows for each $a \in [f(\bar{x}), a_{k+1})$.

Step 4. In Steps 2 and 3, we proved the compactness of $M(\bar{x}, a)$ for $a \in [f(\bar{x}), a_1]$ and $a \in [a_1, a_2]$, respectively. We proceed by using the method of mathematical induction. In Step 3, let the index k run along the set $\{2, \dots, r-1\}$ in numerical order, where, at $k = k_0$, the compactness of $M(\bar{x}, a)$, $a \in [a_{k_0}, a_{k_0+1}]$ is proved by means of the above-shown compactness of $M(\bar{x}, a)$, $a \in [f(\bar{x}), a_{k_0}]$.

Finally, we must show the compactness of $M(\bar{x}, \bar{a})$. Let i always vary in the set A_r . We locally add at x^i a constant $\delta^i > 0$ to f , and we denote the resulting function by \hat{f} . Then the set $\hat{U} = \{x \in R^n | \hat{f}(x) > \bar{a}\}$ is open and $\hat{U} \cap M(\bar{x}, \bar{a})$ is bounded since (V3). Put

$$\hat{M} = M(\bar{x}, \bar{a}) \setminus \left\{ \hat{U} \cup \bigcup_{\{j \in \{1, \dots, p\} | f(x_j) \leq a_{r-1}\} \cup \{0\}} \hat{U}_j \right\}.$$

By (V4), Lemma 3, and Theorem 3, we obtain a C^1 -vector field $\hat{\eta}$ with the flow $\hat{\psi}(x, t)$ satisfying

$$\hat{\psi}(\cdot, \varepsilon)[\hat{M}] \subset M(\bar{x}, \bar{a} - \varepsilon).$$

Using the compactness of $M(\bar{x}, \bar{a} - \varepsilon)$ and an analogous argument as above, it follows that $M(\bar{x}, \bar{a})$ is compact. This completes the proof of (ii).

(iii) We show that Γ is usc-H and lsc-B at each $a \in [f(\bar{x}), \bar{a}]$. Obviously, Γ is lsc-B at $f(\bar{x})$, and, in view of Theorems 4 and 5 (i) (ii), it suffices to prove that $M_0(\bar{x}, a)$ is path-connected for each $a \in (f(\bar{x}), \bar{a}]$. Choose $\hat{a} \in (f(\bar{x}), \bar{a}]$ and $\{\hat{x}^1, \hat{x}^2\} \subset M_0(\bar{x}, \hat{a})$. By Theorem 3 and $M_0(\bar{x}, \hat{a}) \subset M(\bar{x}, \hat{a})$, we obtain that $\{\hat{x}^1, \hat{x}^2\} \subset M(\bar{x}, \max\{f(\hat{x}^1), f(\hat{x}^2)\}) \subset M_0(\bar{x}, \hat{a})$, and, in view of Theorem 1, we obtain the existence of a path in $M_0(\bar{x}, \hat{a})$ connecting \hat{x}^1 and \hat{x}^2 . \square

Remark 6. Theorem 5.2.1 in [8] and Lemma 4 (b) of this paper imply that $M(\bar{x}, a_1)$ and $M(\bar{x}, a_2)$, $a_1 < a_2 \leq \bar{a}$ need not be homotopy-equivalent, since, in the case where $(M_0(\bar{x}, a_2) \setminus M(\bar{x}, a_1)) \cap E(P) \neq \emptyset$, there might exist q such that $\dim H_q(M(\bar{x}, a_1)) \neq \dim H_q(M(\bar{x}, a_2))$. In particular, the continuity of Γ is not sufficient to guarantee homotopy-equivalence of the images of Γ .

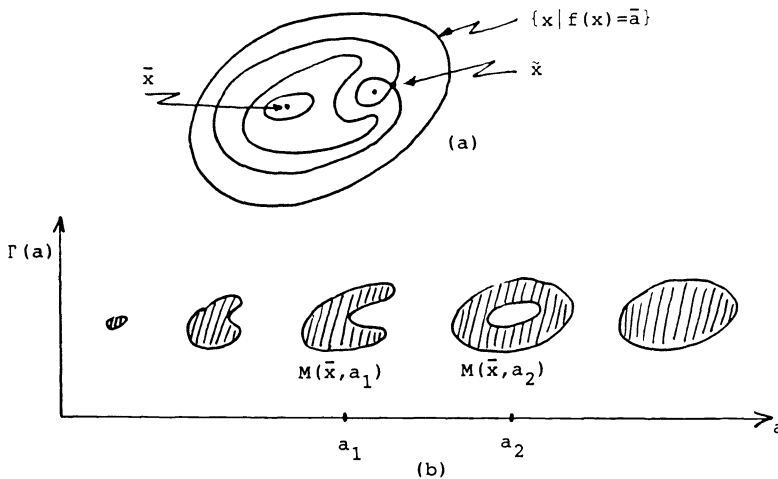


FIG. 5

Remark 7. It is easily seen that assertions (i)–(iii) of Theorem 5 remain true if we assume, instead of s.i. $(x^i) > 1$ in (V3), that (1.3) holds for all $a \in [f(\bar{x}), \bar{a}]$ and s.i. $(x^i) \geq 1$, $i = 1, \dots, p$. In Fig. 5(a), the function f is sketched by means of some level lines, and the three appearing stationary points are strongly stable, having stationary indices 0, 1, and 2. Moreover, it is $\dim H_0(M(\bar{x}, \bar{a}) \cap M^{a_1}) = \dim H_0(M(\bar{x}, a_1))$ and $\dim H_1(M(\bar{x}, a_2)) = \dim H_1(M(\bar{x}, a_1)) + 1$, where $f(\bar{x}) < a_1 < f(\tilde{x}) < a_2 < \bar{a}$, $\tilde{x} \in E(P) \cap M(\bar{x}, \bar{a})$ and s.i. $(\tilde{x}) = 1$ (see Fig. 5(b)).

Finally, we present a corollary in which (V2) is used in a modified form.

COROLLARY 2. *Let MFCQ be satisfied at all $x \in M_0(\bar{x}, \bar{a})$. Then*

(i) $M_0(\bar{x}, \bar{a}) = \bigcup_{a < \bar{a}} M(\bar{x}, a)$,

(ii) *If, in addition, Γ is lsc-B at \bar{a} , then $\text{cl} \bigcup_{a < \bar{a}} M(\bar{x}, a) = M(\bar{x}, \bar{a})$.*

Proof. (i) Obviously, $\bigcup_{a < \bar{a}} M(\bar{x}, a) \subset M_0(\bar{x}, \bar{a})$. Now suppose that $M_0(\bar{x}, \bar{a}) \setminus \bigcup_{a < \bar{a}} M(\bar{x}, a) \neq \emptyset$. Some consideration shows that there exists an $\hat{x} \in M_0(\bar{x}, \bar{a})$ such that, for each neighbourhood $U(\hat{x})$ of \hat{x} , we have that

$$(3.7) \quad U(\hat{x}) \cap \bigcup_{a < \bar{a}} M(\bar{x}, a) \neq \emptyset \quad \text{and}$$

$$(3.8) \quad U(\hat{x}) \cap \left\{ M_0(\bar{x}, \bar{a}) \setminus \bigcup_{a < \bar{a}} M(\bar{x}, a) \right\} \neq \emptyset.$$

By Lemma 1, we obtain a neighbourhood \hat{U} of \hat{x} such that $f(x) < \hat{a} < \bar{a}$ for all $x \in \hat{U}$ and $\hat{U} \cap M^{\hat{a}} = \hat{U} \cap M$ is path-connected. Formula (3.7) implies that $\hat{U} \cap M \subset \bigcup_{a < \bar{a}} M(\bar{x}, a)$, which contradicts (3.8). Assertion (ii) is a consequence of (i) and Theorem 4(i). \square

Acknowledgment. The authors are highly indebted to one of the referees for his precise and constructive criticism.

REFERENCES

- [1] B. BANK, J. GUDDAT, D. KLATTE, B. KUMMER, AND K. TAMMER, *Non-Linear Parametric Optimization*, Akademie-Verlag, Berlin, 1982.
- [2] TH. BRÖCKER AND L. LANDER, *Differential Germs and Catastrophes*, London Math. Soc. Lecture Note Ser., Vol. 17, Cambridge University Press, Cambridge, UK, 1975.
- [3] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [4] J. GAUVIN, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138.
- [5] J. GUDDAT, H. TH. JONGEN, AND J. RUECKMANN, *On stability and stationary points in nonlinear optimization*, J. Austral. Math. Soc. Ser. B, 28 (1986), pp. 36–56.
- [6] H. TH. JONGEN, P. JONKER, AND F. TWILT, *One-parameter families of optimization problems: Equality constraints*, J. Optim. Theory Appl., 48 (1986), pp. 141–161.
- [7] ———, *Critical sets in parametric optimization*, Math. Programming, 34 (1986), pp. 333–353.
- [8] ———, *Nonlinear Optimization in R^n , I. Morse Theory, Chebyshev Approximation*, Peter Lang Verlag, Frankfurt a.M., Bern, New York, 1983.
- [9] H. TH. JONGEN, F. TWILT, AND G. W. WEBER, *Semi-infinite optimization: Structure and stability of the feasible set*, preprint No. 838, Twente University of Technology, Twente, the Netherlands, 1989; J. Optim. Theory Appl., 72 (1992), to appear.
- [10] M. KOJIMA, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 93–138.
- [11] M. KOJIMA AND R. HIRABAYASHI, *Continuous deformation of non-linear programs*, Math. Programming, 21 (1984), pp. 150–198.

- [12] J. MILNOR, *Morse theory*, Annals of Mathematics Studies No. 51, Princeton University Press, Princeton, NJ, 1963.
- [13] S. M. ROBINSON, *Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems*. SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
- [14] R. SCHULZE AND K. TAMMER, *Lokale Konvergenzeigenschaften einer Klasse von Iterationsverfahren der nichtlinearen Optimierung*, Optimization 18 (1987), pp. 677–688.
- [15] E. H. SPANIER, *Algebraic Topology*, McGraw-Hill, New York, 1966.

CONTROL OF A SECOND-ORDER INTEGRO-DIFFERENTIAL EQUATION*

JONG UHN KIM†

Abstract. Reachability for a second-order integro-differential equation is proved. The method is based upon a new kind of unique continuation property. The main significance of the result is that there is no restriction on the size of the memory kernel.

Key words. reachability, integro-differential equation, unique continuation property

AMS(MOS) subject classifications. 35L10, 49E99

Introduction. In this paper we discuss a reachability problem for a second-order integro-differential equation of the following form:

$$(0.1) \quad \begin{aligned} u_{tt}(x, t) - a(t)\Delta u(x, t) + b(t)u_t(x, t) + c(t)u(x, t) \\ - \Delta \int_0^t Q(t, \sigma)u(x, \sigma) d\sigma = 0 \text{ in } \Omega \times (0, T), \end{aligned}$$

$$(0.2) \quad u = g \quad \text{on } \partial\Omega \times (0, T),$$

$$(0.3) \quad u(x, 0) = 0, \quad u_t(x, 0) = 0 \quad \text{in } \Omega.$$

Here Ω is an open bounded subset of R^n with smooth boundary $\partial\Omega$. When $n = 1$, (0.1) reduces to a model equation in linear viscoelasticity. The question of reachability is posed as follows. For given (u_0, u_1) in Ω , is there a boundary control g that can drive the solution of (0.1)–(0.3) to the final state

$$(0.4) \quad u(x, T) = u_0(x), \quad u_t(x, T) = u_1(x) \quad \text{in } \Omega?$$

The purpose of this work is to present an affirmative answer without any size condition on the memory kernel $Q(t, \sigma)$; see Theorem 1.1 below. Let us first review some known results on analogous problems. A reachability problem for a plate equation with a memory was first resolved by Leugering [8]. His main tool was harmonic analysis under the assumption that the space domain is a rectangle and that the memory kernel is in the form of convolution. For a similar equation in a general domain, Lagnese and Lions [5] proved reachability by a different method. Their argument is valid for a general memory kernel including nonconvolution type, but under the assumption that the size of kernel is sufficiently small. Lasiecka [6] also obtained a similar result by a direct operator method with a more general memory kernel that depends both on time and space variables. For a general discussion of analogous problems, the reader is referred to Lions [9]. The common assumption was always that the size of kernel is sufficiently small except in the work of Leugering [8]. In fact, if the size of $Q(t, \sigma)$ in (0.1) is sufficiently small, the above problem can be resolved by the same argument as in [5] and [9]. Here we assume that the memory kernel is independent of space variables in contrast to [6]. By virtue of this assumption, we can employ a fairly simple, but different argument based on a new kind of unique continuation property, which is proved by adapting an idea of Bardos, Lebeau, and Rauch [1]. The technical details are given in the following sections.

* Received by the editors November 26, 1990; accepted for publication (in revised form) September 20, 1991. This work was done at Mathematical Sciences Research Institute, Berkeley, California 94720, under National Science Foundation grant DMS-8505550. The author was also partially supported by Air Force Office of Scientific Research grant 89-0268.

† Department of Mathematics, Virginia Polytechnic Institute, Blacksburg, Virginia 24061.

1. Statement of the main result. We assume that

$$(1.1) \quad a(t) \in C^2([0, \infty)) \text{ and } a(t) \geq a_0, \text{ for all } t, \text{ for some positive constant } a_0,$$

$$(1.2) \quad b(t) \in C^1([0, \infty)),$$

$$(1.3) \quad c(t) \in C([0, \infty)),$$

$$(1.4) \quad Q(\sigma, t) \in C^2([0, \infty) \times [0, \infty)).$$

Let T_0 be a positive number such that

$$\int_0^{T_0} (a(\sigma))^{1/2} d\sigma = \text{diameter of } \Omega.$$

THEOREM 1.1. *Suppose that $T > T_0$. Then, for given $(u_0, u_1) \in L^2(\Omega) \times H^{-1}(\Omega)$, there is a control $g \in L^2(\partial\Omega \times (0, T))$ such that the solution of (0.1)–(0.3) satisfies (0.4).*

By means of a simple transformation of variables, we reduce the above problem to a simpler form.

Let us define $p(t)$ by

$$(1.5) \quad p(t) = \int_0^t (a(\sigma))^{1/2} d\sigma \quad \text{for } t \geq 0,$$

and set

$$(1.6) \quad s = p(t),$$

$$(1.7) \quad v(x, s) = u(x, q(s)),$$

where $q(\cdot)$ is the inverse of $p(\cdot)$, i.e., $t = q(s)$. It is easy to see that (0.1) is equivalent to

$$(1.8) \quad \begin{aligned} & v_{ss}(x, s) - \Delta v(x, s) + a(q(s))^{-1} \{ p''(q(s)) + b(q(s))p'(q(s)) \} v_s(x, s) \\ & + a(q(s))^{-1} c(q(s)) v(x, s) \\ & - \Delta \int_0^s a(q(s))^{-1} Q(q(s), q(\xi)) \frac{dq(\xi)}{d\xi} v(x, \xi) d\xi = 0 \quad \text{in } \Omega \times (0, p(T)), \end{aligned}$$

which can be rewritten as

$$(1.9) \quad \begin{aligned} & v_{ss}(x, s) - \Delta v(x, s) + b_1(s) v_s(x, s) + b_2(s) v(x, s) \\ & - \Delta \int_0^s Q_1(s, \xi) v(x, \xi) d\xi = 0 \quad \text{in } \Omega \times (0, p(T)). \end{aligned}$$

Next we set

$$(1.10) \quad w(x, s) = v(x, s) \exp \left(\int_0^s \frac{1}{2} b_1(\sigma) d\sigma \right).$$

Then (1.9) is equivalent to

$$(1.11) \quad \begin{aligned} & w_{ss}(x, s) - \Delta w(x, s) + \left(b_2(s) - \frac{1}{2} b_1'(s) - \frac{1}{4} b_1^2(s) \right) w(x, s) \\ & - \Delta \int_0^s Q_1(s, \xi) \exp \left(\frac{1}{2} \int_\xi^s b_1(\sigma) d\sigma \right) w(x, \xi) d\xi = 0 \quad \text{in } \Omega \times (0, p(T)), \end{aligned}$$

which we rewrite as

$$(1.12) \quad w_{ss}(x, s) - \Delta w(x, s) + b_3(s)w(x, s) - \Delta \int_0^s Q_2(s, \xi)w(x, \xi) d\xi = 0 \\ \text{in } \Omega \times (0, p(T)).$$

It follows from (1.1)–(1.4) that

$$(1.13) \quad b_3(s) \in C([0, \infty)),$$

$$(1.14) \quad Q_2(s, \xi) \in C^2([0, \infty) \times [0, \infty)).$$

Now we consider a reachability problem for the following equation.

$$(1.15) \quad u_{tt}(x, t) - \Delta u(x, t) + \alpha(t)u(x, t) - \Delta \int_0^t Q(t, \sigma)u(x, \sigma) d\sigma = 0 \quad \text{in } \Omega \times (0, T),$$

$$(1.16) \quad u = g \quad \text{on } \partial\Omega \times (0, T),$$

$$(1.17) \quad u(x, 0) = 0, \quad u_t(x, 0) = 0 \quad \text{in } \Omega,$$

where we assume that

$$(1.18) \quad \alpha(t) \in C([0, \infty)),$$

$$(1.19) \quad Q(t, \sigma) \in C^2([0, \infty) \times [0, \infty)).$$

Our claim is the following.

THEOREM 1.2. *Let T be greater than the diameter of Ω . Then, for given $(u_0, u_1) \in L^2(\Omega) \times H^{-1}(\Omega)$, there is a control $g \in L^2(\partial\Omega \times (0, T))$ that can drive the solution of (1.15)–(1.17) to the final state*

$$(1.20) \quad u(x, T) = u_0(x), \quad u_t(x, T) = u_1(x) \quad \text{in } \Omega.$$

Through the above transformation of variables, it is obvious that Theorem 1.2 implies Theorem 1.1.

The essence of the proof of this theorem is to establish a certain key estimate (Lemma 2.1) of the dual problem. For this, we reduce the dual equation to a simpler equation with a memory as a compact perturbation by solving an associated integral equation. Then, by means of the known estimates for solutions of a wave equation, the proof of the key estimate can be reduced to a unique continuation property of an integro-differential equation. Finally, we resolve this unique continuation property by adapting an idea in the proof of Proposition 6 in [1]. The details are presented in the following sections.

2. Proof of Theorem 1.2. Let us consider the dual problem.

$$(2.1) \quad \phi_{tt} - \Delta \phi + \alpha(t)\phi - \Delta \int_t^T Q(\sigma, t)\phi(x, \sigma) d\sigma = 0 \quad \text{in } \Omega \times (0, T),$$

$$(2.2) \quad \phi = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(2.3) \quad \phi(x, T) = \phi_0(x), \quad \phi_t(x, T) = \phi_1(x) \quad \text{in } \Omega.$$

The crux of the matter is to establish the following key estimate.

LEMMA 2.1. *Let T be greater than the diameter of Ω . For $(\phi_0, \phi_1) \in H_0^1(\Omega) \times L^2(\Omega)$, let ϕ be a unique solution of (2.1)–(2.3) in $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$. Then, it holds that $\partial v / \partial \nu \in L^2(\partial\Omega \times (0, T))$ and*

$$(2.4) \quad \int_0^T \int_{\partial\Omega} \left(\frac{\partial v}{\partial \nu} \right)^2 dx dt \geq M(\|\phi_0\|_{H_0^1(\Omega)}^2 + \|\phi_1\|_{L^2(\Omega)}^2),$$

where M denotes a positive constant independent of ϕ_0 and ϕ_1 , and v is defined by

$$(2.5) \quad v(x, t) = \phi(x, t) + \int_t^T Q(\sigma, t) \phi(x, \sigma) d\sigma.$$

Here $\partial/\partial\nu$ denotes the outward normal derivative on $\partial\Omega$.

We postpone the proof of this and proceed to prove Theorem 1.2. We define a mapping Λ on $H_0^1(\Omega) \times L^2(\Omega)$ as follows. For given $(\phi_0, \phi_1) \in H_0^1(\Omega) \times L^2(\Omega)$, let ϕ be a unique solution of (2.1)–(2.3) in $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$. By choosing $g = \partial v/\partial\nu$, where v is defined by (2.5), we find a unique solution u of (1.15)–(1.17) according to Lemma 3.5 below. Now we define

$$(2.6) \quad \Lambda(\phi_0, \phi_1) = (-u_t(T), u(T)).$$

Then, Λ is a continuous linear mapping from $H_0^1(\Omega) \times L^2(\Omega)$ into $H^{-1}(\Omega) \times L^2(\Omega)$. The continuity of Λ will be obvious in the next section. On account of (2.4) and (3.26), we also have

$$(2.7) \quad \langle \Lambda(\phi_0, \phi_1), (\phi_0, \phi_1) \rangle \geq M(\|\phi_0\|_{H_0^1(\Omega)}^2 + \|\phi_1\|_{L^2(\Omega)}^2),$$

where \langle, \rangle stands for the duality pairing between $H_0^1(\Omega) \times L^2(\Omega)$ and $H^{-1}(\Omega) \times L^2(\Omega)$. Hence, Theorem 1.2 follows from the Lax–Milgram lemma. In the next section, we present some technical preliminaries for the proof of Lemma 2.1.

3. Preliminaries. We consider the following initial boundary value problem.

$$(3.1) \quad v_{tt} - \Delta v = h \quad \text{in } \Omega \times (0, T),$$

$$(3.2) \quad v = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(3.3) \quad v(x, 0) = v_0(x), \quad v_t(x, 0) = v_1(x) \quad \text{in } \Omega.$$

The following facts are well known.

LEMMA 3.1. For $(v_0, v_1) \in H_0^1(\Omega) \times L^2(\Omega)$ and $h \in L^1(0, T; L^2(\Omega))$, there is a unique solution v of (3.1)–(3.3) in $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$. Furthermore, it holds that

$$(3.4) \quad \begin{aligned} & \|v\|_{C([0, T]; H_0^1(\Omega))}^2 + \|v_t\|_{C([0, T]; L^2(\Omega))}^2 + \int_0^T \int_{\partial\Omega} \left(\frac{\partial v}{\partial \nu} \right)^2 dx dt \\ & \leq M(\|v_0\|_{H_0^1(\Omega)}^2 + \|v_1\|_{L^2(\Omega)}^2 + \|h\|_{L^1(0, T; L^2(\Omega))}^2), \end{aligned}$$

where $\partial/\partial\nu$ denotes the outward normal derivative on $\partial\Omega$ and M is a constant independent of v_0 , v_1 , and h .

For the proof, see [9].

LEMMA 3.2. Let $h = 0$ and let T be greater than the diameter of Ω . Then, for $(v_0, v_1) \in H_0^1(\Omega) \times L^2(\Omega)$, it holds that

$$(3.5) \quad \int_0^T \int_{\partial\Omega} \left(\frac{\partial v}{\partial \nu} \right)^2 dx dt \geq M(\|v_0\|_{H_0^1(\Omega)}^2 + \|v_1\|_{L^2(\Omega)}^2),$$

for a positive constant M independent of v_0 and v_1 .

This was originally due to Ho [2] and was later improved in the present form by Komornik [4]. The next lemma will be used to define a weak solution.

LEMMA 3.3. *Let u be a smooth solution of (1.15)–(1.17) with $g \in C_0^\infty(\partial\Omega \times (0, T))$. Then, it holds that*

$$(3.6) \quad \|u\|_{C([0,T];L^2(\Omega))} + \|u_t\|_{C([0,T];H^{-1}(\Omega))} \leq M \|g\|_{L^2(\partial\Omega \times (0,T))},$$

where M is a constant independent of g .

Proof. Fix any $0 < s \leq T$ and let ϕ be a smooth solution of the following dual problem.

$$(3.7) \quad \phi_{tt}(x, t) - \Delta \phi(x, t) + \alpha(t)\phi(x, t) - \Delta \int_t^s Q(\sigma, t)\phi(x, \sigma) d\sigma = 0 \quad \text{in } \Omega \times (0, s),$$

$$(3.8) \quad \phi = 0 \quad \text{on } \partial\Omega \times (0, s),$$

$$(3.9) \quad \phi(x, s) = \phi_0(x), \quad \phi_t(x, s) = \phi_1(x) \quad \text{in } \Omega,$$

where ϕ_0 and ϕ_1 belong to $C_0^\infty(\Omega)$. Multiplying (3.7) by u and integrating over $\Omega \times (0, s)$, we obtain

$$(3.10) \quad \langle \phi_1, u(s) \rangle - \langle \phi_0, u_t(s) \rangle = \int_0^s \int_{\partial\Omega} g \frac{\partial v}{\partial \nu} dx dt,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $L^2(\Omega)$ and

$$(3.11) \quad v(x, t) = \phi(x, t) + \int_t^s Q(\sigma, t)\phi(x, \sigma) d\sigma.$$

This can be inverted to give

$$(3.12) \quad \phi(x, t) = v(x, t) + \int_t^s R(\sigma, t)v(x, \sigma) d\sigma,$$

where $R(\sigma, t)$ is determined from $Q(\sigma, t)$ and $R(\sigma, t) \in C^2([0, \infty) \times [0, \infty))$. It is apparent that v satisfies

$$(3.13) \quad v_0 = v(x, s) = \phi_0, \quad v_1 = v_t(x, s) = \phi_1 - Q(s, s)\phi_0,$$

$$(3.14) \quad v = 0 \quad \text{on } \partial\Omega \times (0, s),$$

$$(3.15) \quad \begin{aligned} &v_{tt}(x, t) - \Delta v(x, t) - R(t, t)v_t(x, t) + (\alpha(t) - 2R_t(t, t) - R_{\sigma\sigma}(t, t))v(x, t) \\ &+ \int_t^s \alpha(t)R(\sigma, t)v(x, \sigma) d\sigma + \int_t^s R_{tt}(\sigma, t)v(x, \sigma) d\sigma = 0 \quad \text{in } \Omega \times (0, s). \end{aligned}$$

It is well known that

$$(3.16) \quad \|v\|_{C([0,s];H_0^1(\Omega))} + \|v_t\|_{C([0,s];L^2(\Omega))} \leq M(\|v_0\|_{H_0^1(\Omega)} + \|v_1\|_{L^2(\Omega)}),$$

where M is a constant depending on T , but independent of v_0, v_1 , and $0 < s \leq T$. By writing (3.15) in the form of (3.1) and modifying the time interval in an obvious manner, we can use (3.4) to derive

$$(3.17) \quad \begin{aligned} \int_0^s \int_{\partial\Omega} \left(\frac{\partial v}{\partial \nu} \right)^2 dx dt &\leq M(\|v_0\|_{H_0^1(\Omega)}^2 + \|v_1\|_{L^2(\Omega)}^2) \\ &\leq M(\|\phi_0\|_{H_0^1(\Omega)}^2 + \|\phi_1\|_{L^2(\Omega)}^2), \end{aligned}$$

where M is independent of ϕ_0, ϕ_1 and $0 < s \leq T$. Now we consider (3.10). By setting $\phi_0 = 0$, we have

$$(3.18) \quad \begin{aligned} |\langle \phi_1, u(s) \rangle| &\leq \|g\|_{L^2(\partial\Omega \times (0,s))} \left\| \frac{\partial v}{\partial \nu} \right\|_{L^2(\partial\Omega \times (0,s))} \\ &\leq M \|g\|_{L^2(\partial\Omega \times (0,T))} \|\phi_1\|_{L^2(\Omega)}, \end{aligned}$$

where M is independent of ϕ_1 , g and $0 < s \leq T$, from which it follows that

$$(3.19) \quad \|u(s)\|_{L^2(\Omega)} \leq M \|g\|_{L^2(\partial\Omega \times (0, T))}$$

for all $0 \leq s \leq T$. Similarly, we set $\phi_1 = 0$ and use (3.17) to obtain

$$(3.20) \quad \|u_t(s)\|_{H^{-1}(\Omega)} \leq M \|g\|_{L^2(\partial\Omega \times (0, T))}$$

for all $0 \leq s \leq T$. The proof of (3.6) is now complete.

By the method of transposition, we can define a weak solution as in [5] and [9].

DEFINITION 3.4. For given $g \in L^2(\partial\Omega \times (0, T))$, a function $u \in C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-1}(\Omega))$ is called a solution of (1.15)–(1.17) if

$$(3.21) \quad \int_0^T \int_{\Omega} u h \, dx \, dt = - \int_0^T \int_{\partial\Omega} g \frac{\partial v}{\partial \nu} \, dx \, dt$$

for every $h \in C_0^\infty(\Omega \times (0, T))$, where

$$(3.22) \quad v(x, t) = \phi(x, t) + \int_t^T Q(\sigma, t) \phi(x, \sigma) \, d\sigma$$

and ϕ is a solution of

$$(3.23) \quad \phi_{tt} - \Delta \phi + \alpha(t) \phi - \Delta \int_t^T Q(\sigma, t) \phi(x, \sigma) \, d\sigma = h \quad \text{in } \Omega \times (0, T),$$

$$(3.24) \quad \phi = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(3.25) \quad \phi(x, T) = 0, \quad \phi_t(x, T) = 0 \quad \text{on } \Omega.$$

It is evident that a smooth solution is also a solution according to the above definition.

LEMMA 3.5. For given $g \in L^2(\partial\Omega \times (0, T))$, there is a unique solution of (1.15)–(1.17).

Proof. We can find a sequence $\{g^m\}_{m=1}^\infty$ in $C_0^\infty(\partial\Omega \times (0, T))$ such that g^m converges to g strongly in $L^2(\partial\Omega \times (0, T))$. For each g^m , there is a smooth solution u^m and $\{u^m\}_{m=1}^\infty$ is strongly convergent in $C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-1}(\Omega))$ by virtue of (3.6). The limit function u is obviously a solution of (1.15)–(1.17). Uniqueness is trivial.

Through a similar procedure, we can also prove the following identity, which was already used in the proof of Theorem 1.2.

LEMMA 3.6. For $g \in L^2(\partial\Omega \times (0, T))$, the above solution also satisfies

$$(3.26) \quad \langle \phi_1, u(T) \rangle - \langle \phi_0, u_t(T) \rangle = \int_0^T \int_{\partial\Omega} g \frac{\partial v}{\partial \nu} \, dx \, dt,$$

where v is defined by (2.5) in terms of $\phi \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$, which is a solution of (2.1)–(2.3). In (3.26), $\langle \cdot, \cdot \rangle$ denotes either the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ or the inner product in $L^2(\Omega)$.

Finally, we present a certain regularity property of solution. Let us set

$$(3.27) \quad w(x, t) = v(x, t) \exp \left(\frac{1}{2} \int_t^T R(\sigma, \sigma) \, d\sigma \right),$$

where v is defined by (2.5) in terms of ϕ , which satisfies (2.1). Then, (3.15) with $s = T$ is equivalent to

$$(3.28) \quad w_{tt} - \Delta w + \gamma(t)w + \int_t^T G(\sigma, t)w(x, \sigma) \, d\sigma = 0 \quad \text{in } \Omega \times (0, T),$$

where $\gamma(t)$ and $G(\sigma, t)$ are expressed in terms of $\alpha(t)$ and $R(\sigma, t)$, and

$$(3.29) \quad \gamma(t) \in C([0, \infty)),$$

$$(3.30) \quad G(\sigma, t) \in C([0, \infty) \times [0, \infty)).$$

LEMMA 3.7. *Let T be greater than the diameter of Ω and let Ω_0 be an open subset of Ω such that $\bar{\Omega}_0 \subset \Omega$. Suppose that $w \in C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-1}(\Omega))$ satisfies (3.28) in the sense of distribution in $\Omega \times (0, T)$. If $w = 0$ in $(\Omega \setminus \Omega_0) \times (0, T)$, then $w \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$.*

Proof. Let $w^\varepsilon = w * \rho_\varepsilon$, where ρ_ε is the Friedrichs mollifier in R^n and the convolution is taken only in the space variables. We take ε so small that $\text{supp } w^\varepsilon \subset \Omega \times [0, T]$. Then, w^ε satisfies (3.28) in $\Omega \times (0, T)$ and $w^\varepsilon \in C^1([0, T]; H_0^2(\Omega))$. We then write

$$(3.31) \quad w^\varepsilon = \psi^\varepsilon + \zeta^\varepsilon,$$

where $\psi^\varepsilon \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ is a solution of

$$(3.32) \quad \psi_{tt}^\varepsilon - \Delta \psi^\varepsilon = -\gamma(t)w^\varepsilon - \int_t^T G(\sigma, t)w^\varepsilon(x, \sigma) d\sigma \quad \text{in } \Omega \times (0, T),$$

$$(3.33) \quad \psi^\varepsilon = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(3.34) \quad \psi^\varepsilon(x, T) = 0, \quad \psi_t^\varepsilon(x, T) = 0 \quad \text{in } \Omega,$$

and ζ^ε is a solution of

$$(3.35) \quad \zeta_{tt}^\varepsilon - \Delta \zeta^\varepsilon = 0 \quad \text{in } \Omega \times (0, T),$$

$$(3.36) \quad \zeta^\varepsilon = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(3.37) \quad \zeta^\varepsilon(x, T) = w^\varepsilon(x, T), \quad \zeta_t^\varepsilon(x, T) = w_t^\varepsilon(x, T) \quad \text{in } \Omega.$$

By virtue of Lemma 3.1 with time reversed, we find that

$$(3.38) \quad \|\psi^\varepsilon\|_{C([0, T]; H_0^1(\Omega))} + \|\psi_t^\varepsilon\|_{C([0, T]; L^2(\Omega))} \leq M \|w\|_{C([0, T]; L^2(\Omega))},$$

$$(3.39) \quad \int_0^T \int_{\partial\Omega} \left(\frac{\partial \psi^\varepsilon}{\partial \nu} \right)^2 dx dt \leq M \|w\|_{C([0, T]; L^2(\Omega))}^2,$$

for a constant M independent of ε and w . In the mean time, since $\text{supp } w^\varepsilon \subset \Omega \times [0, T]$, we have

$$(3.40) \quad \frac{\partial \psi^\varepsilon}{\partial \nu} = -\frac{\partial \zeta^\varepsilon}{\partial \nu} \quad \text{on } \partial\Omega \times (0, T),$$

which, combined with Lemma 3.2 and (3.39), yields

$$(3.41) \quad \|w^\varepsilon(T)\|_{H_0^1(\Omega)}^2 + \|w_t^\varepsilon(T)\|_{L^2(\Omega)}^2 \leq M \|w\|_{C([0, T]; L^2(\Omega))}^2.$$

It now follows that

$$(3.42) \quad \|w^\varepsilon\|_{C([0, T]; H_0^1(\Omega))} + \|w_t^\varepsilon\|_{C([0, T]; L^2(\Omega))} \leq M \|w\|_{C([0, T]; L^2(\Omega))},$$

where M is independent of ε and w . This yields $w \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$.

4. Proof of Lemma 2.1. We shall first prove the following.

LEMMA 4.1. *Let T be greater than the diameter of Ω and let $w \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ be a solution of (3.28) in $\Omega \times (0, T)$ such that $\partial w / \partial \nu = 0$ on $\partial\Omega \times (0, T)$. Then, $w = 0$ in $\Omega \times (0, T)$.*

Proof. Let Ω_1 be a bounded open subset of R^n with smooth boundary such that $\bar{\Omega} \subset \Omega_1$ and let $T > \text{diameter of } \Omega_1 > \text{diameter of } \Omega$. We can extend w to $\Omega_1 \times (0, T)$ such that $w(x, t) = 0$ for $x \in \Omega_1 \setminus \Omega$. Since $w = \partial w / \partial \nu = 0$ on $\partial \Omega \times (0, T)$, $w \in C([0, T]; H_0^1(\Omega_1)) \cap C^1([0, T]; L^2(\Omega_1))$ and w satisfies

$$(4.1) \quad w_{tt} - \Delta w + \gamma(t)w + \int_t^T G(\sigma, t)w(x, \sigma) d\sigma = 0$$

in the sense of distribution in $\Omega_1 \times (0, T)$. We now set

$$(4.2) \quad \mathcal{S} = \{w \in C([0, T]; H_0^1(\Omega_1)) \cap C^1([0, T]; L^2(\Omega_1)) : w \text{ is a solution of (4.1) in } \Omega_1 \times (0, T) \text{ and } w = 0 \text{ in } (\Omega_1 \setminus \Omega) \times (0, T)\}.$$

It is easy to see that \mathcal{S} is a Banach space equipped with the norm of $C([0, T]; H_0^1(\Omega_1)) \cap C^1([0, T]; L^2(\Omega_1))$, which we shall denote by $\|\cdot\|_{\mathcal{S}}$. It will be shown that \mathcal{S} is of finite dimension. Let

$$(4.3) \quad \mathcal{X} = \{w \in \mathcal{S} : \|w\|_{\mathcal{S}} \leq 1\}.$$

If \mathcal{X} is compact, then \mathcal{S} is of finite dimension. For this, we will use the regularity result established in Lemma 3.7 as in [7]. Choose any $w \in \mathcal{X}$ and set

$$(4.4) \quad \eta_i = \frac{\partial w}{\partial x_i}, \quad i = 1, \dots, n.$$

Then, it is evident that Lemma 3.7 can be applied to each η_i . By virtue of (3.42), we find that $\eta_i \in \mathcal{S}$ and

$$(4.5) \quad \|\eta_i\|_{\mathcal{S}} \leq M,$$

where M is a positive constant independent of w , provided $w \in \mathcal{X}$. Hence, it follows that \mathcal{X} is bounded in $C([0, T]; H_0^1(\Omega_1) \cap H^2(\Omega_1)) \cap C^1([0, T]; H_0^1(\Omega_1))$. It is also bounded in $C^2([0, T]; L^2(\Omega_1))$ by (4.1). Therefore, \mathcal{X} is compact. Next choose any $w \in \mathcal{S}$. Since \mathcal{S} is of finite dimension and $\partial/\partial x_1$ is a linear operator from \mathcal{S} into \mathcal{S} , there is an integer $N \geq 1$ such that

$$(4.6) \quad \left(\frac{\partial}{\partial x_1}\right)^N w + \alpha_1 \left(\frac{\partial}{\partial x_1}\right)^{N-1} w + \dots + \alpha_N w = 0 \quad \text{in } \Omega_1 \times (0, T),$$

for some constants $\alpha_1, \dots, \alpha_N$; see [3, p. 191]. We show that $w = 0$ in $\Omega_1 \times (0, T)$. Let us use the notation

$$(4.7) \quad y = x_1 \in R,$$

$$(4.8) \quad z = (x_2, \dots, x_n, t) \in R^n,$$

so that $(x, t) = (y, z)$. Now we choose any $(y_0, z_0) \in \bar{\Omega} \times (0, T)$. Then, there is y_1 such that $(y_1, z_0) \in (\Omega_1 \setminus \bar{\Omega}) \times (0, T)$ and the line segment connecting (y_1, z_0) and (y_0, z_0) is contained in $\Omega_1 \times (0, T)$. There are also positive numbers δ_1 and δ_2 such that the cylinder $I_{\delta_1} \times B_{\delta_2}$ is contained in $\Omega_1 \times (0, T)$ where $I_{\delta_1} = (y_1 - \delta_1, y_0 + \delta_1)$ and $B_{\delta_2} = \{z \in R^n : |z - z_0| < \delta_2\}$. We can further require that $[y_1 - \delta_1, y_1] \times B_{\delta_2}$ is contained in $(\Omega_1 \setminus \bar{\Omega}) \times (0, T)$. Next we choose any $\phi \in C_0^\infty(B_{\delta_2})$ and set

$$(4.9) \quad \xi(y) = \int_{B_{\delta_2}} w \phi dz.$$

Then, $\xi(y) \in L^2(I_{\delta_1})$ and it holds that

$$(4.10) \quad \left(\frac{d}{dy}\right)^N \xi + \alpha_1 \left(\frac{d}{dy}\right)^{N-1} \xi + \dots + \alpha_N \xi = 0$$

in the sense of distribution in I_{δ_1} , which yields that $\xi \in C^\infty(I_{\delta_1})$. In the meantime, $\xi = 0$ in $(y_1 - \delta_1, y_1)$, since $w = 0$ in $(\Omega_1 \setminus \Omega) \times (0, T)$. This implies that $\xi = 0$ in I_{δ_1} . Consequently, $w = 0$ in a neighborhood of (y_0, z_0) . We finally conclude that $w = 0$ in $\Omega_1 \times (0, T)$. Now the proof is complete.

Proof of Lemma 2.1. First, (3.17) implies that $\partial v / \partial \nu \in L^2(\partial\Omega \times (0, T))$. We define w by (3.27). Then, (2.4) is equivalent to

$$(4.11) \quad \int_0^T \int_{\partial\Omega} \left(\frac{\partial w}{\partial \nu} \right)^2 dx dt \cong M(\|w_0\|_{H_0^1(\Omega)}^2 + \|w_1\|_{L^2(\Omega)}^2),$$

where

$$(4.12) \quad w_0(x) = w(x, T), \quad w_1(x) = w_t(x, T) \quad \text{in } \Omega.$$

Assume that (4.11) is false. Then there is a sequence $\{(w_0^m, w_1^m)\}_{m=1}^\infty$ in $H_0^1(\Omega) \times L^2(\Omega)$ such that

$$(4.13) \quad \|w_0^m\|_{H_0^1(\Omega)}^2 + \|w_1^m\|_{L^2(\Omega)}^2 = 1 \quad \text{for all } m,$$

$$(4.14) \quad (w_0^m, w_1^m) \rightarrow (w_0^\infty, w_1^\infty) \text{ weakly in } H_0^1(\Omega) \times L^2(\Omega) \quad \text{as } m \rightarrow \infty,$$

$$(4.15) \quad (w^m, w_t^m) \rightarrow (w^\infty, w_t^\infty) \text{ weak* in } L^\infty(0, T; H_0^1(\Omega)) \\ \times L^\infty(0, T; L^2(\Omega)) \quad \text{as } m \rightarrow \infty$$

$$(4.16) \quad \int_0^T \int_{\partial\Omega} \left(\frac{\partial w^m}{\partial \nu} \right)^2 dx dt \rightarrow 0 \quad \text{as } m \rightarrow \infty.$$

Here each w^m is a solution of (3.28) and

$$(4.17) \quad w^m = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(4.18) \quad w^m(x, T) = w_0^m(x), \quad w_t^m(x, T) = w_1^m(x) \quad \text{in } \Omega,$$

and w^∞ is a solution of (3.28), (4.17) and

$$(4.19) \quad w^\infty(x, T) = w_0^\infty(x), \quad w_t^\infty(x, T) = w_1^\infty(x) \quad \text{in } \Omega.$$

Since $\partial w^\infty / \partial \nu = 0$ on $\partial\Omega \times (0, T)$, which follows from (4.16), we find that $w^\infty = 0$ by virtue of Lemma 4.1. As before, we write

$$(4.20) \quad w^m = \psi^m + \zeta^m,$$

where $\psi^m \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ is a solution of

$$(4.21) \quad \psi_{tt}^m - \Delta \psi^m = -\gamma(t)w^m - \int_t^T G(\sigma, t)w^m(x, \sigma) d\sigma \quad \text{in } \Omega \times (0, T),$$

$$(4.22) \quad \psi^m = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(4.23) \quad \psi^m(x, T) = 0, \quad \psi_t^m(x, T) = 0 \quad \text{in } \Omega,$$

and $\zeta^m \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ is a solution of

$$(4.24) \quad \zeta_{tt}^m - \Delta \zeta^m = 0 \quad \text{in } \Omega \times (0, T),$$

$$(4.25) \quad \zeta^m = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(4.26) \quad \zeta^m(x, T) = w^m(x, T), \quad \zeta_t^m(x, T) = w_t^m(x, T) \quad \text{in } \Omega.$$

Since (4.15) implies that

$$(4.27) \quad w^m \rightarrow 0 \text{ strongly in } L^1(0, T; L^2(\Omega)),$$

it follows from Lemma 3.1 (with time reversed) that

$$(4.28) \quad \int_0^T \int_{\partial\Omega} \left(\frac{\partial \psi^m}{\partial \nu} \right)^2 dx dt \rightarrow 0 \quad \text{as } m \rightarrow \infty.$$

Now we derive from (4.16), (4.20), and (4.28) that

$$(4.29) \quad \int_0^T \int_{\partial\Omega} \left(\frac{\partial \xi^m}{\partial \nu} \right)^2 dx dt \rightarrow 0 \quad \text{as } m \rightarrow \infty.$$

Thus, by Lemma 3.2, we have

$$(4.30) \quad \|w^m(x, T)\|_{H_0^1(\Omega)}^2 + \|w_t^m(x, T)\|_{L^2(\Omega)}^2 \rightarrow 0$$

as $m \rightarrow \infty$, which contradicts (4.13). Now the proof of (4.11) is complete.

Acknowledgments. I would like to acknowledge a very helpful discussion with E. Zuazua in Vorau, Austria. He suggested a transformation in the time variable to reduce the equation to a simpler form. I am also indebted to nice lectures given by V. Komornik at Virginia Tech.

REFERENCES

- [1] C. BARDOS, G. LEBEAU, and J. RAUCH, *Contrôle et stabilisation dans les problèmes hyperboliques*, in *Contrôlabilité exacte Perturbations et Stabilisation de Systèmes Distribués*, Tome 1, Appendice 2, J. L. Lions, ed., Masson, Paris, 1988.
- [2] L. F. HO, *Observabilité frontière de l'équation des ondes*, C. R. Acad. Sci. Paris, Série I, 302 (1986), pp. 443–446.
- [3] K. HOFFMAN AND R. KUNZE, *Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [4] V. KOMORNIK, *Contrôlabilité exacte en un temps minimal*, C. R. Acad. Sci. Paris, Série I, 304 (1987), pp. 223–225.
- [5] J. LAGNESE AND J. L. LIONS, *Modelling, Analysis and Control of Thin Plates*, Masson, Paris, 1988.
- [6] I. LASIECKA, *Controllability of a viscoelastic Kirchhoff plate*, Internat. Ser. Numer. Math., 91 (1989), pp. 237–247.
- [7] I. LASIECKA AND R. TRIGGIANI, *Exact controllability of the Euler–Bernoulli equation with controls in the Dirichlet and Neumann boundary conditions: A nonconservative case*, SIAM J. Control Optimiz., 27 (1989), pp. 330–373.
- [8] G. LEUGERING, *Exact boundary controllability of an integro-differential equation*, Appl. Math. Optim., 15 (1987), pp. 223–250.
- [9] J. L. LIONS, *Contrôlabilité exacte Perturbations et Stabilisation de Systèmes Distribués*, Tomes 1 and 2, Masson, Paris, 1988.

METROPOLIS-TYPE ANNEALING ALGORITHMS FOR GLOBAL OPTIMIZATION IN \mathbb{R}^{d*}

SAUL B. GELFAND[†] AND SANJOY K. MITTER[‡]

Abstract. The convergence of a class of Metropolis-type Markov-chain annealing algorithms for global optimization of a smooth function $U(\cdot)$ on \mathbb{R}^d is established. No prior information is assumed as to what bounded region contains a global minimum. The analysis contained herein is based on writing the Metropolis-type algorithm in the form of a recursive stochastic algorithm $X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k$, where $\{W_k\}$ is a standard white Gaussian sequence, $\{\xi_k\}$ are random variables, and $a_k = A/k$, $b_k = \sqrt{B}/\sqrt{k \log \log k}$ for k large. Convergence results for $\{X_k\}$ are then applied from our previous work [*SIAM Journal on Control and Optimization*, 29 (1991), pp. 999–1018]. Since the analysis of $\{X_k\}$ is based on the asymptotic behavior of the related Langevin-type Markov diffusion annealing algorithm $dY(t) = -\nabla U(Y(t)) dt + c(t) dW(t)$, where $W(\cdot)$ is a standard Wiener process and $c(t) = \sqrt{C}/\sqrt{\log t}$ for t large, this work demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing.

Key words. global optimization, random optimization, simulated annealing, stochastic gradient algorithms, Markov chains

AMS(MOS) subject classifications. 65K10, 90C30, 60J60

1. Introduction. Let $U(\cdot)$ be a real-valued function on some set Σ . The global optimization problem is to find an element of the set $S^* = \{x: U(x) \leq U(y) \text{ for all } y \in \Sigma\}$ (assuming that $S^* \neq \emptyset$). Recently, there has been much interest in the simulated annealing method for global optimization. Annealing algorithms were initially proposed for finite optimization (Σ finite), and later developed for continuous optimization ($\Sigma = \mathbb{R}^d$). An annealing algorithm for finite optimization was first suggested in [17], [2] and is based on simulating a finite-state Metropolis-type Markov chain. The Metropolis algorithm and other related algorithms such as the “heat bath” algorithm, were originally developed as Markov chain sampling methods for sampling from a Gibbs distribution [1]. The asymptotic behavior of finite state Metropolis-type annealing algorithms has been extensively analyzed [3], [5], [9], [12], [14], [21], [24], [25].

A continuous-time annealing algorithm for continuous optimization was first suggested in [10], [13], and is based on simulating a Langevin-type Markov diffusion as follows:

$$(1.1) \quad dY(t) = -\nabla U(Y(t)) dt + c(t) dW(t).$$

Here $U(\cdot)$ is a smooth function on \mathbb{R}^d , $W(\cdot)$ is a standard d -dimensional Wiener process, and $c(\cdot)$ is a positive function with $c(t) \rightarrow 0$ as $t \rightarrow \infty$. In the terminology of simulated annealing algorithms, $U(x)$ is called the energy of state x , and $T(t) = c^2(t)/2$ is called the temperature at time t . Note that for a fixed temperature $T(t) = T$, the resulting Langevin diffusion, like the Metropolis chain, has a Gibbs distribution $\propto \exp(-U(x)/T)$ as its invariant measure. Now (1.1) can be viewed as adding decreasing white Gaussian noise to the continuous time gradient algorithm

$$(1.2) \quad \dot{z}(t) = -\nabla U(z(t)).$$

* Received by the editors July 2, 1990; accepted for publication (in revised form) September 5, 1991. This research was supported by National Science Foundation contract ECS-8910073, Air Force Office of Scientific Research contract 89-0276B, and by Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems).

[†] School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907.

[‡] Department of Electrical Engineering and Computer Science, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

We use (1.1) instead of (1.2) for minimizing $U(\cdot)$ to avoid getting trapped in strictly local minima. The asymptotic behavior of $Y(t)$ as $t \rightarrow \infty$ has been studied in [4], [10], [11], [18]. In [10], [18] convergence results were obtained for a version of (1.1), which was modified to constrain the trajectories to lie in a fixed bounded set (and hence is only applicable to global optimization over a compact subset of \mathbb{R}^d); in [4], [11] results were obtained for global optimization over all of \mathbb{R}^d . Chiang, Hwang, and Sheu's main result from [4] can be roughly stated as follows: If $U(\cdot)$ is suitably behaved and $c^2(t) = C/\log t$ for t large with $C > C_0$ (a constant depending only on $U(\cdot)$), then $Y(t) \rightarrow S^*$ as $t \rightarrow \infty$ in probability.

A discrete-time annealing algorithm for continuous optimization was suggested in [8], [18] and is based on simulating a recursive stochastic algorithm

$$(1.3) \quad X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k.$$

Here $U(\cdot)$ is again a smooth function on \mathbb{R}^d , $\{\xi_k\}$ is a sequence of \mathbb{R}^d -valued random variables, $\{W_k\}$ is a sequence of independent standard d -dimensional Gaussian random variables, and $\{a_k\}, \{b_k\}$ are sequences of positive numbers with $a_k, b_k \rightarrow 0$ as $k \rightarrow \infty$. Algorithm (1.3) could arise from a discretization or numerical integration of the diffusion (1.1) so as to be suitable for implementation on a digital computer; in this case, ξ_k is due to the discretization error. Alternatively, algorithm (1.3) could arise by artificially adding decreasing white Gaussian noise (i.e., the $b_k W_k$ terms) to a stochastic gradient algorithm

$$(1.4) \quad Z_{k+1} = Z_k - a_k(\nabla U(Z_k) + \xi_k),$$

which arises in a variety of optimization problems including adaptive filtering, identification and control; in this case, ξ_k is due to noisy or imprecise measurements of $\nabla U(\cdot)$ (cf. [19]). We again use (1.3) instead of (1.4) for minimizing $U(\cdot)$ to avoid getting trapped in strictly local minima. In the following, we refer to (1.4) and (1.3) as standard and modified stochastic gradient algorithms, respectively. The asymptotic behavior of X_k as $k \rightarrow \infty$ has been studied in [8], [18]. In [18] convergence results were obtained for a version of (1.3), which was modified to constrain the trajectories to lie in a compact set (and hence is only applicable to global optimization over a compact subset of \mathbb{R}^d); in [8] results were obtained for global optimization over all of \mathbb{R}^d . Also, in [18] convergence is obtained essentially only for the case where $\xi_k = 0$; in [8] convergence is obtained for $\{\xi_k\}$ with unbounded variance. This latter fact has important implications when $\nabla U(\cdot)$ is not measured exactly. Our main result from [8] can be roughly stated as follows: If $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ and $b_k^2 = B/k \log \log k$ for k large with $B/A > C_0$ (the same C_0 as above), and $\{X_k\}$ is tight, then $X_k \rightarrow S^*$ as $k \rightarrow \infty$ in probability (conditions are also given in [8] for tightness of $\{X_k\}$). Our analysis in [8] of the asymptotic behavior of X_k as $k \rightarrow \infty$ is based on the behavior of the associated stochastic differential equation (SDE) (1.1). This is analogous to the well-known method of analyzing the asymptotic behavior of Z_k as $k \rightarrow \infty$ based on the behavior of the associated ordinary differential equation (ODE) (1.2) [19], [20].

It has also been suggested that continuous optimization might be performed by simulating a continuous-state Metropolis-type Markov chain [10]. This method has been applied to the restoration of noise corrupted images [16], [23]. In these works, Gaussian random field models are used so that the state space is unbounded. Although some numerical work has been performed with continuous-state Metropolis-type annealing algorithms, there has been very little theoretical analysis, and, furthermore,

the analysis of the continuous-state case does not follow from the finite-state case in a straightforward way (especially for an unbounded state space). The only analysis of which we know is in [16], where a certain asymptotic stability property is established for a related algorithm and a particular cost function that arises in a problem of image restoration.

In this paper, we demonstrate the convergence of a class of continuous-state Metropolis-type Markov-chain annealing algorithms for general cost functions. Our approach is to write such an algorithm in the form of a modified stochastic gradient algorithm (1.3) for suitable choice of ξ_k , and to apply results from [8]. A convergence result is obtained for global optimization over all of \mathbb{R}^d . Some care is necessary to formulate a Metropolis-type Markov chain with appropriate scaling. It turns out that writing the Metropolis-type annealing algorithm in the form (1.3) is more complicated than writing standard variations of gradient algorithms, which use some type of finite-difference estimate of $\nabla U(\cdot)$, in the form (1.4) (cf. [19]). Indeed, to the extent that the Metropolis-type annealing algorithm uses an estimate of $\nabla U(\cdot)$, it does so in a much more subtle manner than a finite-difference approximation, as is seen in the analysis.

Since our convergence results for the Metropolis-type Markov-chain annealing algorithm are ultimately based on the asymptotic behavior of the Langevin-type Markov diffusion annealing algorithm, this paper demonstrates and exploits the close relationship between the Markov chain and diffusion versions of simulated annealing, which is particularly interesting in view of the fact that the development and analysis of these methods has proceeded more or less independently. We note that similar convergence results for other annealing algorithms based on the continuous-state Markov-chain sampling method (such as the “heat bath” method) can be obtained by a procedure similar to that used in this paper.

It is important to note that, although we establish the convergence of the Metropolis-type Markov-chain annealing algorithm by effectively comparing it with the Langevin-type Markov diffusion annealing algorithm, the finite-time behavior of the algorithms may be quite different. Some indication of this arises in the analysis; see Remarks 1 and 2 in § 4.

The paper is organized as follows. In § 2 we discuss appropriately modified versions of tightness and convergence results for modified stochastic gradient algorithms, as given in [8]. In § 3 we present a class of continuous-state Metropolis-type annealing algorithms and state some convergence theorems. In § 4 we prove the convergence theorems of § 3, using the results of § 2.

2. Modified stochastic gradient algorithms. In this section, we give convergence and tightness results for modified stochastic gradient algorithms of the type described in § 1. The algorithms and theorems discussed below are a slight variation on the results of [8] and are appropriate for proving convergence and tightness for a class of continuous state Metropolis-type annealing algorithms (see §§ 3 and 4).

We use the following notation throughout the paper. Let $\nabla U(\cdot)$, $\Delta U(\cdot)$, and $HU(\cdot)$ denote the gradient, Laplacian, and Hessian matrix of $U(\cdot)$, respectively. Let $|\cdot|$, $\langle \cdot, \cdot \rangle$, and \otimes denote Euclidean norm, inner product, and outer product, respectively. For real numbers a and b , let $a \vee b = \text{maximum}\{a, b\}$, $a \wedge b = \text{minimum}\{a, b\}$, $[a]^+ = a \vee 0$, and $[a]^- = a \wedge 0$. For a process $\{X_k\}$ and a function $f(\cdot)$, let $E_{n,x}\{f(X_k)\}$ denote conditional expectation, given $X_n = x$, and let $E_{n_1, x_1; n_2, x_2}\{f(X_k)\}$ denote conditional expectation, given $X_{n_1} = x_1$ and $X_{n_2} = x_2$ (more precisely, these are suitably fixed versions of the conditional expectation). Also, for a measure $\mu(\cdot)$ and a function $f(\cdot)$, let

$\mu(f) = \int f d\mu$. Finally, let $N(m, R)(\cdot)$ denote normal measure with mean m and covariance matrix R , and let I denote the identity matrix.

2.1. Convergence. In this section, we consider the convergence of the discrete-time algorithm

$$(2.1) \quad X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k.$$

Here $U(\cdot)$ is a smooth real-valued function on \mathbb{R}^d , $\{\xi_k\}$ is a sequence of \mathbb{R}^d -valued random variables, $\{W_k\}$ is a sequence of independent standard d -dimensional Gaussian random variables, and

$$a_k = \frac{A}{k}, \quad b_k = \frac{\sqrt{B}}{\sqrt{k \log \log k}}, \quad k \text{ large},$$

where A, B are positive constants.

For $k = 0, 1, \dots$, let $\mathcal{F}_k = \sigma(X_0, W_0, \dots, W_{k-1}, \xi_0, \dots, \xi_{k-1})$. In the following, we consider the following conditions (α, β are constants whose values are specified later).

Condition 1. $U(\cdot)$ is a C^2 function from \mathbb{R}^d to $[0, \infty)$ such that

$$\begin{aligned} \lim_{|x| \rightarrow \infty} \frac{|\nabla U(x)|}{|x|} &> 0, \\ \lim_{|x| \rightarrow \infty} \left\langle \frac{\nabla U(x)}{|\nabla U(x)|}, \frac{x}{|x|} \right\rangle &= 1, \\ \inf_x (|\nabla U(x)|^2 - \Delta U(x)) &> -\infty. \end{aligned}$$

Condition 2. For $\varepsilon > 0$, let

$$d\pi^\varepsilon(x) = \frac{1}{Z^\varepsilon} \exp\left(-\frac{2U(x)}{\varepsilon^2}\right) dx, \quad Z^\varepsilon = \int \exp\left(-\frac{2U(x)}{\varepsilon^2}\right) dx < \infty.$$

π^ε has a weak limit π as $\varepsilon \rightarrow 0$.

Condition 3. Let K be a compact subset of \mathbb{R}^d . Then there exists $L, k_0 \geq 0$ such that, for every $k \geq k_0$,

$$(2.2a) \quad E\{|\xi_k|^2 | \mathcal{F}_k\} \leq L a_k^\alpha, \quad \forall X_k \in K, \quad \text{with probability one (w.p.1),}$$

$$(2.2b) \quad |E\{\xi_k | \mathcal{F}_k\}| \leq L a_k^\beta, \quad \forall X_k \in K, \quad \text{w.p.1.}$$

W_k is independent of \mathcal{F}_k .

We note that π concentrates on S^* , the global minima of $U(\cdot)$. The existence of π and a simple characterization in terms of $HU(\cdot)$ is discussed in [15].

In [4] and [8], it was shown that there exists a constant C_0 , which plays a critical role in the convergence of (1.1) and (1.3), respectively (in [4] C_0 was denoted by c_0). C_0 has an interpretation in terms of the action functional for the perturbed dynamical systems

$$(2.3) \quad dY^\varepsilon(t) = -\nabla U(Y^\varepsilon(t)) dt + \varepsilon dW(t).$$

Now, for $\phi(\cdot)$ an absolutely continuous function on \mathbb{R}^d , the (normalized) action functional for (2.3) is given by

$$I(t, x, y) = \inf_{\substack{\phi(0)=x \\ \phi(t)=y}} \frac{1}{2} \int_0^t |\dot{\phi}(s) + \nabla U(\phi(s))|^2 ds.$$

According to [4],

$$C_0 = \frac{3}{2} \sup_{x, y \in S_0} (V(x, y) - 2U(y)),$$

where $V(x, y) = \lim_{t \rightarrow \infty} I(t, x, y)$, and S_0 is the set of all the stationary points of $U(\cdot)$, i.e., $S_0 = \{x: \nabla U(x) = 0\}$; see [4] for a further discussion of C_0 , including some examples.

Let $K_1 \subset \mathbb{R}^d$, and let $\{X_k^x\}$ denote the solution of (2.1) with $X_0 = x$. We say that $\{X_k^x: k \geq 0, x \in K_1\}$ is tight if, given $\varepsilon > 0$, there exists a compact $K_2 \subset \mathbb{R}^d$ such that $P_{0,x}\{X_k \in K_2\} > 1 - \varepsilon$ for all $k \geq 0$ and $x \in K_1$. Below is our theorem on the convergence of X_k as $k \rightarrow \infty$.

THEOREM 1. *Assume that Conditions 1-3 hold with $\alpha > -1$ and $\beta > 0$. Let $\{X_k\}$ be given by (2.1), and assume that $\{X_k^x: k \geq 0, x \in K\}$ is tight for K a compact set. Then, for $B/A > C_0$ and any bounded continuous function $f(\cdot)$ on \mathbb{R}^d ,*

$$(2.4) \quad \lim_{k \rightarrow \infty} E_{0,x}\{f(X_k)\} = \pi(f)$$

uniformly for x in a compact set.

Note that since π concentrates on S^* , under the conditions of Theorem 1, we have that $X_k \rightarrow S^*$ as $k \rightarrow \infty$ in probability.

Theorem 1 is the same as [8, Thm. 2], except there we assumed that (2.2) was valid for all $k \geq 0$. However, examination of the proof of [8, Thm. 2] shows that we actually established that

$$(2.5) \quad \lim_{k \rightarrow \infty} E_{0,x;k_0,x_0}\{f(X_k)\} = \pi(f)$$

uniformly for x_0 in a compact set and all x , only assuming that (2.2) is valid for all $k \geq k_0$. It is easy to show that (2.4) follows from (2.5) and the assumption that $\{X_k^x: k \geq 0, x \in K\}$ is tight.

2.2. Tightness. In this section, we consider the tightness of the discrete-time algorithm

$$(2.6) \quad X_{k+1} = X_k - a_k(\psi_k(X_k) + \eta_k) + b_k \sigma_k(X_k) W_k.$$

Here $\{\psi_k(\cdot)\}$ are Borel functions from \mathbb{R}^d to \mathbb{R}^d , $\{\sigma_k(\cdot)\}$ are Borel functions from \mathbb{R}^d to \mathbb{R} , $\{\eta_k\}$ is a sequence of \mathbb{R}^d -valued random variables, and $\{W_k\}, \{a_k\}, \{b_k\}$ are as in § 2.1. Below, we give sufficient conditions for the tightness of $\{X_k^x: k \geq 0, x \in K\}$, where K is a compact subset of \mathbb{R}^d . Note that algorithm (2.6) is somewhat more general than algorithm (2.1). We consider this more general algorithm because it is sometimes convenient to write an algorithm in the form (2.6) (with $\psi_k(x) \neq \nabla U(x)$ for some x, k) to verify tightness, and then to write the algorithm in the form (2.1) to verify convergence. We give an example of this situation when we consider continuous-state Metropolis-type annealing algorithms in §§ 3 and 4.

Let $\mathcal{G}_k = \sigma(X_0, W_0, \dots, W_{k-1}, \eta_0, \dots, \eta_{k-1})$. We consider the following conditions ($\alpha, \beta, \gamma_1, \gamma_2$ are constants whose values are specified later).

Condition 4. Let K be a compact subset of \mathbb{R}^d . Then

$$\sup_{k; x \in K} |\psi_k(x)| < \infty,$$

$$\overline{\lim}_{k, |x| \rightarrow \infty} \frac{|\psi_k(x)|}{|x|} a_k^{\gamma_1} < \infty,$$

$$\lim_{k, |x| \rightarrow \infty} \frac{|\psi_k(x)|}{|x|} a_k^{\gamma_2} > 0,$$

$$\lim_{k, |x| \rightarrow \infty} \left\langle \frac{\psi_k(x)}{|\psi_k(x)|}, \frac{x}{|x|} \right\rangle > 0.$$

Condition 5. Let K be a compact subset of \mathbb{R}^d . Then

$$\sup_{k; x \in K} |\sigma_k(x)| < \infty, \quad \overline{\lim}_{k, |x| \rightarrow \infty} \frac{|\sigma_k(x)|}{|x|} < \infty.$$

Condition 6. There exists $L \geq 0$ such that

$$(2.7a) \quad E\{|\eta_k|^2 | \mathcal{G}_k\} \leq La_k^\alpha (|X_k|^2 + 1) \quad \text{w.p.1,}$$

$$(2.7b) \quad |E\{\eta_k | \mathcal{G}_k\}| \leq La_k^\beta (|X_k| + 1) \quad \text{w.p.1.}$$

W_k is independent of \mathcal{G}_k .

THEOREM 2. Assume that Conditions 4–6 hold with $\alpha > -1$, $\beta > 0$, and $0 \leq \gamma_2 \leq \gamma_1 < \frac{1}{2}$. Let $\{X_k\}$ be given by (2.6), and let K be a compact subset of \mathbb{R}^d . Then $\{X_k^x: k \geq 0, x \in K\}$ is a tight family of random variables.

Theorem 2 is proved similarly to [8, Thm. 3], where we assumed that $\sigma_k(\cdot) = 1$ and did not allow the bounds in (2.7) to be state-dependent. The extension to the present case is straightforward.

3. Metropolis-type annealing algorithms. In this section, we review the finite-state Metropolis-type Markov-chain annealing algorithm, generalize it to an arbitrary state space, and then specialize it to a class of algorithms for which the results in § 2 can be applied to establish convergence.

The finite-state Metropolis-type annealing algorithm may be described as follows [12]. Assume that the state space Σ is finite set. Let $U(\cdot)$ be a real-valued function on Σ (the “energy” function) and $\{T_k\}$ be a sequence of strictly positive numbers (the “temperature” sequence). Let $q(i, j)$ be a stationary transition probability from i to j , for $i, j \in \Sigma$. The one-step transition probability at time k for the finite-state Metropolis-type annealing chain $\{X_k\}$ is given by

$$(3.1) \quad P\{X_{k+1} = j | X_k = i\} = q(i, j) s_k(i, j), \quad j \neq i,$$

$$P\{X_{k+1} = i | X_k = i\} = 1 - \sum_{j \neq i} q(i, j) s_k(i, j),$$

where

$$(3.2) \quad s_k(i, j) = \exp \left(- \frac{[U(j) - U(i)]^+}{T_k} \right).$$

This nonstationary Markov chain may be interpreted (and simulated) in the following manner. Given the current state $X_k = i$, generate a candidate state $\tilde{X}_k = j$ with probability $q(i, j)$. Set the next state $X_{k+1} = j$ if $s_k(i, j) > \theta_k$, where θ_k is an independent random variable uniformly distributed on the interval $[0, 1]$; otherwise, set $X_{k+1} = i$. Suppose that the stochastic transition matrix $Q = [q(i, j)]$ is symmetric, i.e., $q(i, j) = q(j, i)$, and the temperature T_k is fixed at a constant $T > 0$. Then it is easy to show that the resulting stationary Markov chain has a Gibbs invariant measure with mass $\propto \exp(-U(i)/T)$. Furthermore, if the chain is recurrent, then the chain, in fact, has a unique Gibbs

invariant probability measure, and the transition probabilities converge to the Gibbs probabilities as $k \rightarrow \infty$ for all initial states. Of course, if a finite-state Markov chain is irreducible, then it is recurrent. There has been much work on the convergence and asymptotic behavior of the nonstationary annealing chain when $T_k \rightarrow 0$ [3], [5], [9], [12], [14], [21], [24], [25].

We next generalize the finite-state Metropolis-type annealing algorithm (3.1), (3.2) to a general state space. In the formulation and analysis of general state space Markov chains, it is usually assumed that the state space Σ is a σ -finite measure space, say (Σ, Λ, μ) (see [22, Chap. 1] for a thorough discussion of general state space Markov chains). Let $U(\cdot)$ be a real-valued measurable function on such a Σ , and let $\{T_k\}$ be as above. Let $q(x, y)$ be a stationary transition probability density with respect to μ from x to y , for $x, y \in \Sigma$. The one-step transition probability at time k for the general state Metropolis-type annealing chain $\{X_k\}$ is given by

$$(3.3) \quad P\{X_{k+1} \in A | X_k = x\} = \int_A q(x, y) s_k(x, y) d\mu(y) + r_k(x) 1_A(x),$$

where

$$(3.4) \quad s_k(x, y) = \exp\left(-\frac{[U(y) - U(x)]^+}{T_k}\right)$$

($r_k(x)$ gives the appropriate normalization, i.e., $r_k(x) = 1 - \int q(x, y) s_k(x, y) d\mu(y)$). Note that if μ does not have an atom at x , then $r_k(x)$ is the self transition probability starting at state x at time k . Also, note that (3.3), (3.4) reduces to (3.1), (3.2) when the state space is finite and μ is counting measure. The general state chain may be interpreted (and simulated) similarly to the finite-state chain: here $q(x, y)$ is a conditional probability density for generating a candidate state $\tilde{X}_k = y$, given the current state $X_k = x$. Suppose that the stochastic transition function $Q(x, A) = \int_A q(x, y) d\mu(y)$ is symmetric, i.e., $q(x, y) = q(y, x)$, and the temperature T_k is fixed at a constant $T > 0$. Then it is easy to show that the resulting stationary Markov chain has a Gibbs invariant measure with density (with respect to μ) $\propto \exp(-U(x)/T)$. Furthermore, if this measure is finite and the chain is μ -recurrent,¹ then the chain, in fact, has a unique Gibbs invariant probability measure, and the transition probability measure converges to the Gibbs measure (in the total variation norm) as $k \rightarrow \infty$ for all initial states [22, Thm. 7.1, p. 30]. It is known that if a chain is μ -irreducible² and satisfies a certain condition due to Doeblin [6, Hyp. (D), p. 192], then it is μ -recurrent. In [7, Chap. 3], we use this theory to give some sufficient conditions for the ergodicity of general state Metropolis-type Markov chains when Σ is a compact metric space and μ is a finite Borel measure. However, there has been almost no work on the convergence and asymptotic behavior of the nonstationary annealing chain when $T_k \rightarrow 0$, although, when Σ is a compact metric space, we would expect the behavior to be similar to when Σ is finite.

We next specialize the general state Metropolis-type annealing algorithm (3.3), (3.4) to a d -dimensional Euclidean state space. This is the most important case and the one that has seen application [16], [23]. Actually, the Metropolis-type annealing chain that we consider is not exactly a specialization of the general state chain described above. Motivated by our desire to show convergence of the chain by writing it in the

¹ If, for every $x \in \Sigma$ and $A \in \Lambda$ such that $\mu(A) > 0$, $P_{0,x} \bigcup_{k=1}^{\infty} \{X_k \in A\} = 1$, then $\{X_k^x\}$ is μ -recurrent.

² If, for every $x \in \Sigma$ and $A \in \Lambda$ such that $\mu(A) > 0$, $P_{0,x} \bigcup_{k=1}^{\infty} \{X_k \in A\} > 0$, then $\{X_k^x\}$ is μ -irreducible.

form of the modified stochastic gradient algorithm (2.1), we are led to choosing a nonstationary Gaussian transition density

$$(3.5) \quad q_k(x, y) = \frac{1}{(2\pi b_k^2 \sigma_k^2(x))^{d/2}} \exp\left(-\frac{1}{2} \frac{|y-x|^2}{b_k^2 \sigma_k^2(x)}\right),$$

and a state-dependent temperature sequence

$$(3.6) \quad T_k(x) = \frac{b_k^2 \sigma_k^2(x)}{2a_k} \left(= \frac{\text{const } \sigma_k^2(x)}{\log \log k} \right),$$

where

$$(3.7) \quad \sigma_k(x) = (\delta_k |x|) \vee 1, \quad \delta_k \downarrow 0.$$

To understand these choices, suppose that x lies in some fixed compact set. Then, for k large enough,

$$(3.8) \quad q_k(x, y) = \frac{1}{(2\pi b_k^2)^{d/2}} \exp\left(-\frac{1}{2} \frac{|y-x|^2}{b_k^2}\right)$$

and

$$(3.9) \quad T_k(x) = T_k = \frac{b_k^2}{2a_k}.$$

The choice of the transition density (3.8) is clear, given that we want to write the chain in the form (2.1). The choice of the temperature schedule (3.9) is also clear if we view (2.1) as a sampled version of the associated diffusion (1.1) with sampling intervals a_k and sampling times $t_k = \sum_{n=0}^{k-1} a_n$, since then we should have the corresponding sampled temperatures $T(t_k) = c^2(t_k)/2$. Indeed, it is straightforward to check that, if $C = B/A$, then

$$T_k = \frac{b_k^2}{2a_k} \sim \frac{c^2(t_k)}{2} = T(t_k) \quad \text{as } k \rightarrow \infty$$

(recall that $a_k = A/k$, $b_k^2 = B/k \log \log k$, and $c^2(t) = C/\log t$ for large k, t). Finally, the reason that the $|x|$ dependence is needed in $\sigma_k(x)$, and hence both (3.5), (3.6), is that to establish tightness of the annealing chain by writing the chain in the form of (2.6), we need a condition similar to the following:

$$|\psi_k(x)| \geq \text{const } |x|, \quad |x| \text{ large, } k \text{ fixed,}$$

for suitable choice of $\psi_k(\cdot)$. In other words, the annealing chain must generate a drift (toward the origin) at least proportional to the distance from the origin. This discussion leads us to the following continuous-state Metropolis-type Markov-chain annealing algorithm and convergence result. To establish convergence, we must assume, along with Conditions 1 and 2, the following condition.

Condition 7. It holds that

$$\inf_{\delta > 0} \overline{\lim}_{|x| \rightarrow \infty} \sup_{|y-x| < \delta |x|} |HU(y)| \frac{|x|^2}{U(x)} < \infty.$$

This condition is satisfied if, for example, $U(x) \sim \text{const } |x|^p$ and $HU(x) = O(|x|^{p-2})$ as $|x| \rightarrow \infty$, for some $p \geq 2$.

Metropolis-Type Annealing Algorithm 1. Let $\{X_k\}$ be a Markov chain with one-step transition probability at time k given by

$$(3.10) \quad P\{X_{k+1} \in A \mid X_k = x\} = \int_A s_k(x, y) dN(x, b_k^2 \sigma_k^2(x) I)(y) + r_k(x) 1_A(x),$$

where

$$(3.11) \quad \sigma_k(x) = (a_k^\gamma |x|) \vee 1,$$

$$(3.12) \quad s_k(x, y) = \exp \left(-\frac{2a_k [U(y) - U(x)]^+}{b_k^2 \sigma_k^2(x)} \right),$$

and $\gamma > 0$ ($r_k(x)$ gives the correct normalization).

THEOREM 3. Assume that Conditions 1, 2, and 7 hold, and also that

$$(3.13) \quad \overline{\lim}_{|x| \rightarrow \infty} \frac{|\nabla U(x)|}{|x|} < \infty.$$

Let $\{X_k\}$ be the Markov chain with transition probability given by (3.10)–(3.12) and with $0 < \gamma < \frac{1}{4}$. Then, for $B/A > C_0$ and any bounded continuous function $f(\cdot)$ on \mathbb{R}^d ,

$$(3.14) \quad \lim_{k \rightarrow \infty} E_{0,x}\{f(X_k)\} = \pi(f)$$

uniformly for x in a compact set.

The proof of Theorem 3 is in § 4.1. Observe that the conditions of Theorem 3 are satisfied if, for example, $\nabla U(x) \sim \text{const } x$ and $HU(x) = O(1)$ as $|x| \rightarrow \infty$. We can allow for faster growth in $\nabla U(x)$ by using a suitable modification of (3.12).

Metropolis-Type Annealing Algorithm 2. Let $\{X_k\}$ be a Markov chain with one-step transition probability at time k given by

$$(3.15) \quad P\{X_{k+1} \in A \mid X_k = x\} = \int_A s_k(x, y) dN(x, b_k^2 \sigma_k^2(x) I)(y) + r_k(x) 1_A(x),$$

where

$$(3.16) \quad \sigma_k(x) = (a_k^\gamma |x|) \vee 1,$$

$$(3.17) \quad s_k(x, y) = \exp \left(-\frac{2a_k [U(y) - U(x)]^+}{b_k^2 \sigma_k^2(x)} \right) \quad \text{if } U(x) \leq \frac{|x|^2 + 1}{a_k^\gamma}$$

$$= \exp \left(-\frac{2a_k [|y|^2 - |x|^2]^+}{b_k^2 \sigma_k^2(x)} \right) \quad \text{if } U(x) > \frac{|x|^2 + 1}{a_k^\gamma},$$

and $\gamma > 0$ ($r_k(x)$ gives the correct normalization). Note that if K is any fixed compact, $X_{k=x} \in K$, and k is very large, then (3.17) and (3.12) coincide. Note also that (3.17), like (3.12), only uses measurements of $U(\cdot)$ (and not $\nabla U(\cdot)$).

THEOREM 4. Assume that Conditions 1, 2, and 7 hold, and also that

$$(3.18) \quad \overline{\lim}_{|x| \rightarrow \infty} |\nabla U(x)| \frac{|x|}{U(x)} < \infty.$$

Let $\{X_k\}$ be the Markov chain with transition probability given by (3.15)–(3.17) and with $0 < \gamma < \frac{1}{8}$. Then, for $B/A > C_0$ and any bounded continuous function $f(\cdot)$ on \mathbb{R}^d ,

$$(3.19) \quad \lim_{k \rightarrow \infty} E_{0,x}\{f(X_k)\} = \pi(f)$$

uniformly for x in a compact set.

The proof of Theorem 4 is in § 4.2. Observe that the conditions of Theorem 4 are satisfied if, for example, $\nabla U(x) \sim \text{const } |x|^{p-2}x$ and $HU(x) = O(|x|^{p-2})$ as $|x| \rightarrow \infty$, for some $p \geq 2$.

4. Proofs of Theorems 3 and 4. In the following, c_1, c_2, \dots denotes positive constants whose value may change from proof to proof. We need the following lemma.

LEMMA 1. Assume that $V(\cdot)$ is a C^2 function from \mathbb{R}^d to \mathbb{R} . Let

$$s(x, y) = \exp(-\lambda[V(y) - V(x)]^+)$$

and

$$\hat{s}(x, y) = \exp(-\lambda[\langle \nabla V(x), y - x \rangle]^+),$$

where $\lambda > 0$. Then

$$|s(x, y) - \hat{s}(x, y)| \leq \lambda \sup_{\varepsilon \in (0,1)} |HV(x + \varepsilon(y - x))| |y - x|^2$$

for all $x, y \in \mathbb{R}^d$.

Proof. Let

$$f(x, y) = V(y) - V(x) - \langle \nabla V(x), y - x \rangle.$$

Then, by the second-order Taylor theorem,

$$(4.1) \quad |f(x, y)| \leq \sup_{\varepsilon \in (0,1)} |HV(x + \varepsilon(y - x))| |y - x|^2.$$

By separately considering the four cases corresponding to the possible signs of $V(y) - V(x)$ and $\langle \nabla V(x), y - x \rangle$, it can be shown that

$$(4.2) \quad |s(x, y) - \hat{s}(x, y)| \leq 1 - \exp(-\lambda|f(x, y)|) \leq \lambda|f(x, y)|.$$

Combining (4.1) and (4.2) completes the proof. \square

4.1. Proof of Theorem 3. We write

$$(4.3) \quad X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k$$

(this defines ξ_k) and apply Theorem 1 to show that, if $\{X_k^x; k \geq 0, x \in K\}$ is tight for K compact, then (3.14) is true. We further let $\psi(x) = \nabla U(x)$, write

$$(4.4) \quad X_{k+1} = X_k - a_k(\psi(X_k) + \eta_k) + b_k \sigma_k(X_k) W_k$$

(this defines η_k), and apply Theorem 2 to show that $\{X_k^x; k \geq 0, x \in K\}$ is, in fact, tight for K compact, and that (3.14) is, in fact, true.

We first show that we can find a version of $\{X_k\}$ in the form

$$(4.5) \quad X_{k+1} = X_k + b_k \sigma_k(X_k) \zeta_k W_k.$$

To do this, we inductively define the sequence $\{W_k, \zeta_k\}$ of random variables as follows. Assume that $X_0, W_0, \dots, W_{k-1}, \zeta_0, \dots, \zeta_{k-1}$ have been defined. Let W_k be a standard d -dimensional Gaussian random variable independent of $X_0, W_0, \dots, W_{k-1}, \zeta_0, \dots, \zeta_{k-1}$, and let ζ_k be a $\{0, 1\}$ -valued random variable with

$$(4.6) \quad P\{\zeta_k = 1 | X_0, W_0, \dots, W_k, \zeta_0, \dots, \zeta_{k-1}\} = s_k(X_k, X_k + b_k \sigma_k(X_k) W_k).$$

Using (4.6), it is easy to check that (4.5) is a Markov chain that has transition probability given by (3.10)–(3.12). Hence (4.5) is indeed a version of $\{X_k\}$, and we henceforth always deal with this version.

Now, comparing (4.3) and (4.4) with (4.5), we have that

$$(4.7) \quad \xi_k = -\nabla U(X_k) + \frac{b_k}{a_k} (1 - \sigma_k(X_k) \zeta_k) W_k$$

and

$$(4.8) \quad \eta_k = -\psi(X_k) + \frac{b_k}{a_k} \sigma_k(X_k) (1 - \zeta_k) W_k.$$

Furthermore, it is easy to show that W_k is independent of \mathcal{F}_k and \mathcal{G}_k , and also that $P\{\xi_k \in \cdot \mid \mathcal{F}_k\} = P\{\xi_k \in \cdot \mid X_k\}$ and $P\{\eta_k \in \cdot \mid \mathcal{G}_k\} = P\{\eta_k \in \cdot \mid X_k\}$. We use these facts below.

The following lemmas give the crucial estimates for $E\{|\xi_k|^2 \mid \mathcal{F}_k\}$, $E\{\xi_k \mid \mathcal{F}_k\}$, $E\{|\eta_k|^2 \mid \mathcal{G}_k\}$, and $E\{\eta_k \mid \mathcal{G}_k\}$.

LEMMA 2. *Let K be a compact subset of \mathbb{R}^d . There exists $L, k_0 \geq 0$ such that, for every $k \geq k_0$,*

- (a) $|E\{\xi_k \mid \mathcal{F}_k\}| \leq L(a_k/b_k)$ for all $X_k \in K$, w.p.1;
- (b) $E\{|\xi_k|^2 \mid \mathcal{F}_k\} \leq L(b_k/a_k)$ for all $X_k \in K$, w.p.1.

LEMMA 3. *There exists $L \geq 0$ such that*

- (a) $|E\{\eta_k \mid \mathcal{G}_k\}| \leq L(a_k^{1-2\gamma}/b_k)(|X_k|+1)$ w.p.1;
- (b) $E\{|\eta_k|^2 \mid \mathcal{G}_k\} \leq L(b_k/a_k^{1+\gamma})(|X_k|^2+1)$ w.p.1.

Assume that Lemmas 2 and 3 are true. Then Condition 3 is satisfied with $\alpha = -\frac{1}{2} > -1$ and $0 < \beta < \frac{1}{2}$. Conditions 4–6 are satisfied for $\alpha = -\frac{1}{2} - \gamma > -1$, $0 < \beta < \frac{1}{2} - 2\gamma$, and $\gamma_1 = \gamma_2 = 0$ (recall that we assume that $0 < \gamma < \frac{1}{4}$). Hence Theorems 1 and 2 apply, and Theorem 3 follows. It remains to prove Lemmas 2 and 3. We use the following claim.

CLAIM. *Let $u \in \mathbb{R}^d$ with $|u| = 1$. Then*

- (a) $\int_{0 \leq \langle u, w \rangle \leq \delta} dN(0, I)(w) = O(\delta)$;
- (b) $\int_{0 \leq \langle u, w \rangle \leq \delta} w dN(0, I)(w) = O(\delta^2)$;
- (c) $\int_{0 \leq \langle u, w \rangle \leq \delta} w \otimes w dN(0, I)(w) = O(\delta)$.

Proof. Let $u_1 = u$, and extend u_1 to an orthonormal basis $\{u_1, \dots, u_d\}$ for \mathbb{R}^d . Then, by changing variables (rotation) and using the mean value theorem, we obtain that

$$\begin{aligned}
 (a) \quad & \int_{0 \leq \langle u, w \rangle \leq \delta} dN(0, I)(w) = \int_0^\delta \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta); \\
 (b) \quad & \int_{0 \leq \langle u, w \rangle \leq \delta} w dN(0, I)(w) = u_1 \int_0^\delta v \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta^2); \\
 (c) \quad & \int_{0 \leq \langle u, w \rangle \leq \delta} w \otimes w dN(0, I)(w) = u_1 \otimes u_1 \int_0^\delta v^2 \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv \\
 & \quad + \sum_{i=2}^d u_i \otimes u_i \int_0^\delta \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{v^2}{2}\right) dv = O(\delta).
 \end{aligned}$$

□

Proof of Lemma 2(a). Since K is compact and $a_k^\gamma \rightarrow 0$, we can choose k_0 such that $a_k^\gamma |X_k| \leq 1$ (and so $\sigma_k(X_k) = 1$) for all $X_k \in K$ and $k \geq k_0$. Hence, using (4.7) and the fact that $P\{\xi_k \in \cdot \mid \mathcal{F}_k\} = P\{\xi_k \in \cdot \mid X_k\}$ and W_k is independent of X_k , we have, for $k \geq k_0$ and $X_k \in K$ (w.p.1), that

$$\begin{aligned}
 (4.9) \quad & E\{\xi_k \mid \mathcal{F}_k\} = E\{\xi_k \mid X_k\} \\
 & = -\nabla U(X_k) + \frac{b_k}{a_k} E\{(1 - \zeta_k) W_k \mid X_k\} \\
 & = -\nabla U(X_k) - \frac{b_k}{a_k} E\{W_k E\{\zeta_k \mid X_k, W_k\} \mid X_k\} \\
 & = -\nabla U(X_k) - \frac{b_k}{a_k} E\{W_k P\{\zeta_k = 1 \mid X_k, W_k\} \mid X_k\} \\
 & = -\nabla U(X_k) - \frac{b_k}{a_k} E\{W_k P\{\zeta_k = 1 \mid X_k, W_k\}\}.
 \end{aligned}$$

Henceforth, we assume that $k \geq k_0$ and condition on $X_k = x \in K$. Then, using (4.6), we obtain that

$$(4.10) \quad E\{\xi_k | X_k = x\} = -\nabla U(x) - \frac{b_k}{a_k} \int w s_k(x, x + b_k w) dN(0, I)(w).$$

Let

$$(4.11) \quad \hat{s}_k(x, y) = \exp\left(-\frac{2a_k}{b_k^2} [\langle \nabla U(x), y - x \rangle]^+\right)$$

and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Using the fact that $HU(\cdot) = O(1)$ on a compact, we obtain that, for any fixed $\delta > 0$,

$$\sup_{\varepsilon \in (0,1)} |HU(x + \varepsilon(y - x))| \leq c_1,$$

for all $|y - x| < \delta$. Hence, using Lemma 1,

$$(4.12) \quad |\tilde{s}_k(x, y)| \leq c_2 \frac{a_k}{b_k^2} |y - x|^2, \quad |y - x| < \delta.$$

Of course,

$$(4.13) \quad |\tilde{s}_k(x, y)| \leq 1.$$

Using (4.12), (4.13), and a standard estimate for the tail probability of a Gaussian random variable, we obtain, for $i \geq 0$, that

$$\begin{aligned} & \int |w|^i |\tilde{s}_k(x, x + b_k w)| dN(0, I)(w) \\ & \leq \int_{|w| \leq \delta/b_k} |w|^i |\tilde{s}_k(x, x + b_k w)| dN(0, I)(w) \\ (4.14) \quad & + \int_{|w| > \delta/b_k} |w|^i |\tilde{s}_k(x, x + b_k w)| dN(0, I)(w) \\ & \leq c_3 a_k + c_3 \exp\left(-\frac{c_4}{b_k^2}\right) \\ & = O(a_k). \end{aligned}$$

Now, expanding (4.10) and using (4.14) gives

$$\begin{aligned} E\{\xi_k | X_k = x\} &= -\nabla U(x) - \frac{b_k}{a_k} \int w \hat{s}_k(x, x + b_k w) dN(0, I)(w) \\ &\quad - \frac{b_k}{a_k} \int w \tilde{s}_k(x, x + b_k w) dN(0, I)(w) \\ (4.15) \quad &= -\nabla U(x) - \frac{b_k}{a_k} \int w \hat{s}_k(x, x + b_k w) dN(0, I)(w) \\ &\quad + O(b_k) \end{aligned}$$

$$\begin{aligned} (4.16) \quad &= -\nabla U(x) - \frac{b_k}{a_k} \int_{\langle \nabla U(x), w \rangle \leq 0} w dN(0, I)(w) \\ &\quad - \frac{b_k}{a_k} \int_{\langle \nabla U(x), w \rangle > 0} w \exp\left(-\frac{2a_k}{b_k} \langle \nabla U(x), w \rangle\right) dN(0, I)(w) \\ &\quad + O(b_k). \end{aligned}$$

Clearly,

$$(4.17) \quad E\{\xi_k | X_k = x\} = O(b_k)$$

for x such that $\nabla U(x) = 0$. Henceforth, we assume that $\nabla U(x) \neq 0$. Let $\nabla \hat{U}(x) = \nabla U(x)/|\nabla U(x)|$. Completing the square in the second integral in (4.16), we obtain that

$$(4.18) \quad \begin{aligned} E\{\xi_k | X_k = x\} &= -\nabla U(x) - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w \rangle \leq 0} w dN(0, I)(w) \\ &\quad - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w \rangle \geq 0} w \exp\left(2\left(\frac{a_k}{b_k}\right)^2 |\nabla U(x)|^2\right) dN\left(\frac{2a_k}{b_k} \nabla U(x), I\right)(w) \\ &\quad + O(b_k). \end{aligned}$$

Now $\nabla U(x) = O(1)$, and so

$$(4.19) \quad \exp\left(2\left(\frac{a_k}{b_k}\right)^2 |\nabla U(x)|^2\right) = 1 + O\left(\left(\frac{a_k}{b_k}\right)^2\right).$$

Substituting (4.19) into (4.18), using $\nabla U(x) = O(1)$ and $a_k/b_k = O(1)$, and changing variables from $w + 2(a_k/b_k)\nabla U(x)$ to w , gives

$$(4.20) \quad \begin{aligned} E\{\xi_k | X_k = x\} &= -\nabla U(x) - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w \rangle \leq 0} w dN(0, I)(w) \\ &\quad - \frac{b_k}{a_k} \int_{\langle \nabla \hat{U}(x), w \rangle \geq O(a_k/b_k)} w dN(0, I)(w) \\ &\quad + 2\nabla U(x) \int_{\langle \nabla \hat{U}(x), w \rangle \geq O(a_k/b_k)} dN(0, I)(w) \\ &\quad + O\left(\frac{a_k}{b_k}\right) + O(b_k) \\ &= \frac{b_k}{a_k} \int_{0 \leq \langle \nabla \hat{U}(x), w \rangle \leq O(a_k/b_k)} w dN(0, I)(w) \\ &\quad - 2\nabla U(x) \int_{0 \leq \langle \nabla \hat{U}(x), w \rangle \leq O(a_k/b_k)} dN(0, I)(w) \\ &\quad + O\left(\frac{a_k}{b_k}\right). \end{aligned}$$

Hence, by (a) and (b) of the claim, and by again using $\nabla U(x) = O(1)$, we have that

$$(4.21) \quad E\{\xi_k | X_k = x\} = O\left(\frac{a_k}{b_k}\right).$$

Combining (4.17) and (4.21) completes the proof of Lemma 2(a). \square

Proof of Lemma 2(b). As in the proof of Lemma 2(a), choose k_0 such that $a_k^\gamma |X_k| \leq 1$ (and so $\sigma_k(X_k) = 1$) for all $X_k \in K$ and $k \geq k_0$. Hence, using (4.7) and the fact that

$P\{\xi_k \in \cdot | \mathcal{F}_k\} = P\{\xi_k \in \cdot | X_k\}$, W_k is independent of X_k , and $\nabla U(\cdot) = O(1)$ on a compact, we have, for $k \geq k_0$ and $X_k \in K$ (w.p.1), that

$$\begin{aligned}
 E\{\xi_k \otimes \xi_k | \mathcal{F}_k\} &= E\{\xi_k \otimes \xi_k | X_k\} \\
 &= \left(\frac{b_k}{a_k}\right)^2 E\{((1 - \zeta_k) W_k) \otimes ((1 - \zeta_k) W_k) | X_k\} + e_k(X_k) \\
 (4.22) \quad &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 E\{W_k \otimes W_k E\{\zeta_k | X_k, W_k\} | X_k\} + e_k(X_k) \\
 &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 E\{W_k \otimes W_k P\{\zeta_k = 1 | X_k, W_k\} | X_k\} + e_k(X_k) \\
 &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 E_{W_k}\{W_k \otimes W_k P\{\zeta_k = 1 | X_k, W_k\}\} + e_k(X_k),
 \end{aligned}$$

where

$$\begin{aligned}
 e_k(X_k) &= O\left(\frac{b_k}{a_k} |\nabla U(X_k)| + |\nabla U(X_k)|^2\right) \\
 &= O\left(\frac{b_k}{a_k}\right).
 \end{aligned}$$

Henceforth, we assume that $k \geq k_0$ and condition on $X_k = x \in K$. Then, using (4.6), we obtain that

$$\begin{aligned}
 E\{\xi_k \otimes \xi_k | X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w s_k(x, x + b_k w) dN(0, I)(w) \\
 (4.23) \quad &+ O\left(\frac{b_k}{a_k}\right).
 \end{aligned}$$

Let $\hat{s}_k(x, y)$ be given by (4.11) and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Then, expanding (4.23) and using (4.14), gives

$$\begin{aligned}
 E\{\xi_k \otimes \xi_k | X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w \hat{s}_k(x, x + b_k w) dN(0, I)(w) \\
 &\quad - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w \tilde{s}_k(x, x + b_k w) dN(0, I)(w) \\
 &\quad + O\left(\frac{b_k}{a_k}\right) \\
 (4.24) \quad &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int w \otimes w \hat{s}_k(x, x + b_k w) dN(0, I)(w) \\
 &\quad + O\left(\frac{b_k^2}{a_k}\right) + O\left(\frac{b_k}{a_k}\right)
 \end{aligned}$$

$$\begin{aligned}
 (4.25) \quad &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla U(x), w \rangle \leq 0} w \otimes w dN(0, I)(w) \\
 &\quad - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla U(x), w \rangle > 0} w \otimes w \\
 &\quad \cdot \exp\left(-\frac{2a_k}{b_k} \langle \nabla U(x), w \rangle\right) dN(0, I)(w) \\
 &\quad + O\left(\frac{b_k}{a_k}\right).
 \end{aligned}$$

Clearly,

$$(4.26) \quad E\{\xi_k \otimes \xi_k \mid X_k = x\} = O\left(\frac{b_k}{a_k}\right)$$

for x such that $\nabla U(x) = 0$. Henceforth, we assume that $\nabla U(x) \neq 0$. Let $\nabla \hat{U}(x) = \nabla U(x)/|\nabla U(x)|$. Completing the square in the second integral in (4.25), we obtain that

$$(4.27) \quad \begin{aligned} E\{\xi_k \otimes \xi_k \mid X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \hat{U}(x), w \rangle \leq 0} w \otimes w dN(0, I)(w) \\ &\quad - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla \hat{U}(x), w \rangle \geq 0} w \otimes w \\ &\quad \cdot \exp\left(2\left(\frac{a_k}{b_k}\right)^2 (|\nabla U(x)|)^2\right) dN\left(-\frac{2a_k}{b_k} \nabla U(x), I\right)(w) \\ &\quad + O\left(\frac{b_k}{a_k}\right). \end{aligned}$$

Substituting (4.19) into (4.27), using $\nabla U(x) = O(1)$ and $a_k/b_k = O(1)$, and changing variables from $w + 2(a_k/b_k)\nabla U(x)$ to w gives

$$(4.28) \quad \begin{aligned} E\{\xi_k \otimes \xi_k \mid X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 I - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla \hat{U}(x), w \rangle \leq 0} w \otimes w dN(0, I)(w) \\ &\quad - \left(\frac{b_k}{a_k}\right)^2 \int_{\langle \nabla \hat{U}(x), w \rangle \geq O(a_k/b_k)} w \otimes w dN(0, I)(w) \\ &\quad + O(1) + O\left(\frac{b_k}{a_k}\right) \\ &= \left(\frac{b_k}{a_k}\right)^2 \int_{0 \leq \langle \nabla \hat{U}(x), w \rangle \leq O(a_k/b_k)} w \otimes w dN(0, I)(w) \\ &\quad + O\left(\frac{b_k}{a_k}\right). \end{aligned}$$

Hence, by part (c) of the Claim,

$$(4.29) \quad E\{\xi_k \otimes \xi_k \mid X_k = x\} = O\left(\frac{b_k}{a_k}\right).$$

Combining (4.26) and (4.29) and using $|\xi_k|^2 \leq |\xi_k \otimes \xi_k|$ completes the proof of Lemma 2(b). \square

Proof of Lemma 3. Using (4.8) and the fact that $P\{\eta_k \in \cdot \mid \mathcal{G}_k\} = P\{\eta_k \in \cdot \mid X_k\}$, W_k is independent of X_k , and $\psi(x) = \nabla U(x) = O(|x| + 1)$, we get, similarly to (4.9) and (4.22), that

$$E\{\eta_k \mid \mathcal{G}_k\} = -\psi(X_k) - \frac{b_k}{a_k} \sigma_k(X_k) E\{W_k P\{\zeta_k = 1 \mid X_k, W_k\}\}$$

and

$$E\{\eta_k \otimes \eta_k | \mathcal{G}_k\} = \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) I \\ - \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) E_{W_k} \{W_k \otimes W_k P\{\zeta_k = 1 | X_k, W_k\}\} + e_k(X_k),$$

where

$$e_k(X_k) = O\left(\frac{b_k}{a_k} \sigma_k(X_k) |\psi(X_k)| + |\psi(X_k)|^2\right) \\ = O\left(\frac{b_k}{a_k} \sigma_k(X_k) (|X_k| + 1) + |X_k|^2\right).$$

Henceforth, we condition on $X_k = x$ and assume, for simplicity, that $|x| \geq 1$ and $a_k \leq 1$.

Let

$$(4.30) \quad \hat{s}_k(x, y) = \exp\left(-\frac{2a_k [\langle \nabla U(x), y - x \rangle]^+}{b_k^2 \sigma_k^2(x)}\right)$$

and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Using Condition 7, we obtain, for some $\delta > 0$, that

$$\sup_{\varepsilon \in (0, 1)} |HU(x + \varepsilon(y - x))| \leq c_1 \left(\frac{U(x)}{|x|^2} + 1\right)$$

for all $|y - x| < \delta|x|$. By assumption, however, $\nabla U(x) = O(|x|)$, and so, by the mean value theorem, $U(x) = O(|x|^2)$. Hence, using Lemma 1, we obtain that

$$|\tilde{s}_k(x, y)| \leq c_2 \frac{a_k}{b_k^2} \frac{|y - x|^2}{\sigma_k^2(x)}, \quad |y - x| < \delta \sigma_k(x),$$

and, similarly to the derivation of (4.14), we obtain that

$$(4.31) \quad \int |w|^i \tilde{s}_k(x, x + b_k \sigma_k(x) w) dN(0, I)(w) = O(a_k).$$

Next, using (4.31), we obtain, similarly to the derivation of (4.15) and (4.24), that

$$E\{\eta_k | X_k = x\} = -\psi(x) - \frac{b_k}{a_k} \sigma_k(x) \int w \hat{s}_k(x, x + b_k \sigma_k(x) w) dN(0, I)(w) \\ + O(b_k \sigma_k(x))$$

and

$$E\{\eta_k \otimes \eta_k | X_k = x\} = \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) I \\ - \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) \int w \otimes w \hat{s}_k(x, x + b_k \sigma_k(x) w) dN(0, I)(w) \\ + O\left(\frac{b_k^2}{a_k} \sigma_k^2(x)\right) + O\left(\frac{b_k}{a_k} \sigma_k(x) |x| + |x|^2\right).$$

At this point, we separately consider the cases where $a_k^\gamma |x| \leq 1$ and > 1 ($\sigma_k(x) = 1$ and $a_k^\gamma |x|$, respectively). Proceeding as in the proof of Lemma 2 and using $\psi(x) = O(|x|)$ and $a_k^{1-\gamma}/b_k = O(1)$, we can show that

$$\begin{aligned} E\{\eta_k | X_k = x\} &= O\left(\frac{a_k^{1-2\gamma}}{b_k}\right), & a_k^\gamma |x| \leq 1, \\ &= O\left(\frac{a_k^{1-\gamma}}{b_k} |x|\right), & a_k^\gamma |x| > 1, \end{aligned}$$

and

$$\begin{aligned} E\{\eta_k \otimes \eta_k | X_k = x\} &= O\left(\frac{b_k}{a_k^{1+\gamma}}\right), & a_k^\gamma |x| \leq 1, \\ &= O\left(\frac{b_k}{a_k^{1-\gamma}} |x|^2\right), & a_k^\gamma |x| > 1. \end{aligned}$$

Combining the two cases completes the proof of the lemma. \square

Remark 1. In Fig. 1 we demonstrate the type of approximations used in the proof of Theorem 3. In Fig. 1(a) we show the transition density $p_k(x, y)$ for the Markov chain with transition probability given by (3.10)–(3.12); in Figure 1(b) we show the transition density $p'_k(x, y)$ for the same Markov chain but using $\hat{s}_k(x, y)$ (4.30) in place of $s_k(x, y)$ (3.12); and in Fig. 1(c) we show the transition density $p''_k(x, y)$ for the Markov chain of (2.1) with $\xi_k = 0$. Note that the densities in Figs. 1(a) and 1(b) contain impulsive components associated with the positive probability of no transition. All three densities are “close” for sufficiently large k and x in a compact set, and this is the basis of the proof. However, for small k , the transition densities can be quite different. In particular, it is seen that the Metropolis-type algorithm takes a less “local” point of view than the gradient-based algorithms.

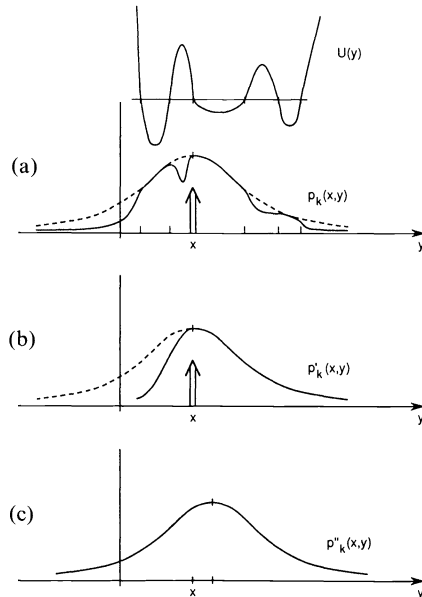


FIG. 1. Three transition probability densities.

Remark 2. The Metropolis-type Markov-chain annealing algorithms use only measurements of $U(\cdot)$ (and not $\nabla U(\cdot)$). Another class of algorithms that use only measurements of $U(\cdot)$ could be based on a finite-difference approximation $D_k U(\cdot)$ of $\nabla U(\cdot)$,

$$(4.32) \quad X_{k+1} = X_k - a_k D_k U(X_k) + b_k W_k.$$

Suppose that $D_k U(\cdot)$ is a random direction forward finite-difference approximation; i.e., suppose that θ_k is an independent random vector uniformly distributed on the d -one-dimensional unit sphere, and

$$D_k U(x) = \frac{U(x + h_k \theta_k) - U(x)}{h_k} \theta_k$$

($\{h_k\}$ is a sequence of nonzero numbers with $h_k \rightarrow 0$). If we write (4.32) in the form (2.1), then, by analysis similar to [19, pp. 58–60], it can be shown that ξ_k is *bounded* for X_k in a compact. However, when we write (3.10)–(3.12) in the form (2.1), the best estimate we can obtain suggests that ξ_k is *unbounded* for X_k in a compact (see Lemma 2(b), and note that $b_k/a_k \rightarrow \infty$). Hence the Metropolis-type approximation appears to be much farther away from an exact gradient-based algorithm than a finite-difference approximation.

4.2. Proof of Theorem 4. We write

$$X_{k+1} = X_k - a_k (\nabla U(X_k) + \xi_k) + b_k W_k$$

(this defines ξ_k) and apply Theorem 1 to show that, if $\{X_k^x: k \geq 0, x \in K\}$ is tight for K compact, then (3.19) is true. We further let

$$\begin{aligned} \psi_k(x) &= \nabla U(x) \quad \text{if } U(x) \leq \frac{|x|^2 + 1}{a_k^\gamma} \\ &= 2x \quad \text{if } U(x) > \frac{|x|^2 + 1}{a_k^\gamma}, \end{aligned}$$

write

$$X_{k+1} = X_k - a_k (\psi_k(X_k) + \eta_k) + b_k \sigma_k(X_k) W_k$$

(this defines η_k), and apply Theorem 2 to show that $\{X_k^x: k \geq 0, x \in K\}$ is, in fact, tight for K compact and (3.19) is, in fact, true.

The following lemmas give the crucial estimates for $E\{|\xi_k|^2 | \mathcal{F}_k\}$, $|E\{\xi_k | \mathcal{F}_k\}|$, $E\{|\eta_k|^2 | \mathcal{G}_k\}$, and $|E\{\eta_k | \mathcal{G}_k\}|$ (compare with Lemmas 2 and 3).

LEMMA 4. *Let K be a compact subset of \mathbb{R}^d . Then there exists $L, k_0 \geq 0$ such that, for every $k \geq k_0$,*

- (a) $|E\{\xi_k | \mathcal{F}_k\}| \leq L(a_k/b_k)$ for all $X_k \in K$, w.p.1;
- (b) $E\{|\xi_k|^2 | \mathcal{F}_k\} \leq L(b_k/a_k)$ for all $X_k \in K$, w.p.1.

LEMMA 5. *There exists $L \geq 0$ such that*

- (a) $|E\{\eta_k | \mathcal{G}_k\}| \leq L(a_k^{1-4\gamma}/b_k)(|X_k| + 1)$ w.p.1;
- (b) $E\{|\eta_k|^2 | \mathcal{G}_k\} \leq L(b_k/a_k^{1+2\gamma})(|X_k|^2 + 1)$ w.p.1.

Assume that Lemmas 4 and 5 are true. Then Condition 3 is satisfied with $\alpha = -\frac{1}{2} > -1$ and $0 < \beta < \frac{1}{2}$. Conditions 4–6 are satisfied with $\alpha = -\frac{1}{2} - 2\gamma > -1$, $0 < \beta < \frac{1}{2} - 4\gamma$, $\gamma_1 = \gamma$, and $\gamma_2 = 0$ (recall that we assume that $0 < \gamma < \frac{1}{8}$). We note that the second relation in Condition 4 is verified with $\gamma_1 = \gamma$ by considering the two cases where $U(x)$ is $<$ or $\geq (|x|^2 + 1)/a_k^\gamma$ and applying (3.18); in fact, we obtain that $\psi_k(x) = O(|x|/a_k^\gamma)$ as $|x| \rightarrow \infty$ uniformly for all k . Hence Theorems 1 and 2 apply, and Theorem 4 follows. It remains to prove Lemmas 4 and 5.

Proof of Lemma 4. Since K is compact and $a_k^\gamma \rightarrow 0$, we can choose k_0 such that $U(X_k) \leq (|X_k|^2 + 1)/a_k^\gamma$ for all $X_k \in K$ and $k \geq k_0$. Hence the proof of Lemma 4 is the same as that of Lemma 2. \square

Proof of Lemma 5. Using $\psi_k(x) = O(|x|/a_k^\gamma + 1)$ (see discussion following the statement of Lemmas 4 and 5), we get, similarly to the proof of Lemma 3, that

$$E\{\eta_k | \mathcal{G}_k\} = -\psi_k(X_k) - \frac{b_k}{a_k} \sigma_k(X_k) E_{W_k}\{W_k P\{\zeta_k = 1 | X_k, W_k\}\}$$

and

$$\begin{aligned} E\{\eta_k \otimes \eta_k | \mathcal{G}_k\} &= \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) I \\ &\quad - \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(X_k) E_{W_k}\{W_k \otimes W_k P\{\zeta_k = 1 | X_k, W_k\}\} + e_k(X_k), \end{aligned}$$

where

$$\begin{aligned} e_k(X_k) &= O\left(\frac{b_k}{a_k} \sigma_k(X_k) |\psi_k(X_k)| + |\psi_k(X_k)|^2\right) \\ &= O\left(\frac{b_k}{a_k} \sigma_k(X_k) \left(\frac{|X_k|}{a_k^\gamma} + 1\right) + \frac{|X_k|^2}{a_k^{2\gamma}}\right). \end{aligned}$$

Henceforth, we condition on $X_k = x$ and assume for simplicity that $|x| \geq 1$ and $a_k \leq 1$.

Let

$$\begin{aligned} \hat{s}_k(x, y) &= \exp\left(-\frac{2a_k}{b_k^2} \frac{[\langle \nabla U(x), y - x \rangle]^+}{\sigma_k^2(x)}\right) \quad \text{if } U(x) \leq \frac{|x|^2 + 1}{a_k^\gamma} \\ &= \exp\left(-\frac{2a_k}{b_k^2} \frac{[\langle 2x, y - x \rangle]^+}{\sigma_k^2(x)}\right) \quad \text{if } U(x) > \frac{|x|^2 + 1}{a_k^\gamma} \end{aligned}$$

and $\tilde{s}_k(x, y) = s_k(x, y) - \hat{s}_k(x, y)$. Now if a C^2 function $V(\cdot)$ satisfies Condition 7, then, for some $\delta > 0$,

$$\sup_{\varepsilon \in (0, 1)} |HV(x + \varepsilon(y - x))| \leq c_1 \left(\frac{V(x)}{|x|^2} + 1\right)$$

for all $|y - x| < \delta|x|$; so this inequality holds when $V(z) = U(z)$ and when $V(z) = |z|^2$. Hence, by considering the two cases when $U(x)$ is \leq or $>$ $(|x|^2 + 1)/a_k^\gamma$ and using Lemma 1, we obtain that

$$|\tilde{s}_k(x, y)| \leq c_2 \frac{a_k^{1-\gamma}}{b_k^2} \frac{|y - x|^2}{\sigma_k^2(x)}, \quad |y - x| < \delta \sigma_k(x),$$

and, similarly to the derivation of (4.14), we obtain that

$$(4.33) \quad \int |w|^i \tilde{s}_k(x, x + b_k \sigma_k(x) w) dN(0, I)(w) = O(a_k^{1-\gamma}).$$

Next, using (4.33), we obtain, similarly to the derivation of (4.15) and (4.24), that

$$\begin{aligned} E\{\eta_k | X_k = x\} &= -\psi_k(x) - \frac{b_k}{a_k} \sigma_k(x) \int w \hat{s}_k(x, x + b_k \sigma_k(x) w) dN(0, I)(w) \\ &\quad + O\left(\frac{b_k}{a_k^\gamma} \sigma_k(x)\right) \end{aligned}$$

and

$$\begin{aligned} E\{\eta_k \otimes \eta_k | X_k = x\} &= \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) I \\ &\quad - \left(\frac{b_k}{a_k}\right)^2 \sigma_k^2(x) \int w \otimes w \hat{s}_k(x, x + b_k \sigma_k(x) w) dN(0, I)(w) \\ &\quad + O\left(\frac{b_k^2}{a_k^{1+\gamma}} \sigma_k^2(x)\right) + O\left(\frac{b_k}{a_k^{1+\gamma}} \sigma_k(x) |x| + \frac{|x|^2}{a_k^{2\gamma}}\right). \end{aligned}$$

We now separately consider the cases where $a_k^\gamma |x| \leq 1$ and > 1 ($\sigma_k(x) = 1$ and $a_k^\gamma |x|$, respectively). Proceeding as in the proof of Lemma 2 and using $\psi_k(x) = O(|x|/a_k^\gamma)$ and $a_k^{1-2\gamma}/b_k = O(1)$, we can show that

$$\begin{aligned} E\{\eta_k | X_k = x\} &= O\left(\frac{a_k^{1-4\gamma}}{b_k}\right), & a_k^\gamma |x| \leq 1, \\ &= O\left(\frac{a_k^{1-3\gamma}}{b_k} |x|\right), & a_k^\gamma |x| > 1 \end{aligned}$$

and

$$\begin{aligned} E\{\eta_k \otimes \eta_k | X_k = x\} &= O\left(\frac{b_k}{a_k^{1+2\gamma}}\right), & a_k^\gamma |x| \leq 1, \\ &= O\left(\frac{b_k}{a_k} |x|^2\right), & a_k^\gamma |x| > 1. \end{aligned}$$

Combining the two cases completes the proof of the lemma. \square

Acknowledgments. The authors thank the referees for a careful reading of the manuscript, which uncovered an important technical problem, and for suggesting its solution.

REFERENCES

- [1] K. BINDER, *Monte Carlo Methods in Statistical Physics*, Springer-Verlag, Berlin, 1978.
- [2] V. CERNY, *A thermodynamical approach to the travelling salesman problem*, J. Optim. Theory Appl., 45 (1985), pp. 41–51.
- [3] T. S. CHIANG AND Y. CHOW, *On the convergence rate of annealing processes*, SIAM J. Control Optim., 26 (1988), pp. 1455–1470.
- [4] T. S. CHIANG, C. R. HWANG, AND S. J. SHEU, *Diffusion for global optimization in \mathbb{R}^n* , SIAM J. Control Optim., 25 (1987), pp. 737–752.
- [5] D. P. CONNORS AND P. R. KUMAR, *Simulated annealing type Markov chains and their order balance equations*, SIAM J. Control Optim., 27 (1989), pp. 1440–1461.
- [6] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
- [7] S. B. GELFAND, *Analysis of simulated annealing type algorithms*, Ph.D. thesis, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Report No. LIDS-TH-1668, Cambridge, MA, 1987.
- [8] S. B. GELFAND AND S. K. MITTER, *Recursive stochastic algorithms for global optimization in \mathbb{R}^d* , SIAM J. Control Optim., 29 (1991), pp. 999–1018.
- [9] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intelligence, PAMI-6 (1984), pp. 721–741.
- [10] S. GEMAN AND C. R. HWANG, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1986), pp. 1031–1043.
- [11] B. GIDAS, *Global optimization via the Langevin equation*, in Proc. IEEE Conf. on Decision and Control, Fort Lauderdale, FL, 1985, pp. 774–778.

- [12] ———, *Nonstationary Markov chains and convergence of the annealing algorithm*, J. Statist. Phys., 39 (1985), pp. 73–131.
- [13] U. GRENENDER, *Tutorial in Pattern Theory*, Division of Applied Mathematics, Brown University, Providence, RI, 1984.
- [14] B. HAJEK, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.
- [15] C. R. HWANG, *Laplaces method revisited: Weak convergence of probability measures*, Ann. Probab., 8 (1980), pp. 1177–1182.
- [16] F. C. JENG AND J. W. WOODS, *Simulated annealing in compound Gaussian random fields*, IEEE Trans. Inform. Theory, IT-36 (1990), pp. 94–107.
- [17] S. KIRKPATRICK, C. D. GELATT, AND M. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 621–680.
- [18] H. J. KUSHNER, *Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo*, SIAM J. Appl. Math., 47 (1987), pp. 169–185.
- [19] H. J. KUSHNER AND D. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer, Berlin, 1978.
- [20] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.
- [21] D. MITRA, F. ROMEO, AND A. SANGIOVANNI-VINCENTELLI, *Convergence and finite-time behavior of simulated annealing*, Adv. Appl. Probab., 18 (1986), pp. 747–771.
- [22] S. OREY, *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold, London, 1971.
- [23] T. SIMCHONY, R. CHELLAPA, AND Z. LICHTENSTEIN, *Relaxation algorithms for MAP estimation of gray-level images with multiplicative noise*, IEEE Trans. Inform. Theory, IT-36 (1990), pp. 608–614.
- [24] J. TSITSIKLIS, *A survey of large time asymptotics of simulated annealing algorithms*, Report No. LIDS-P-1623, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1986.
- [25] ———, *Markov chains with rare transitions and simulated annealing*, Math. Oper. Res., 14 (1989), pp. 70–90.

SINGULAR PERTURBATIONS IN MANUFACTURING*

H. METE SONER†

Abstract. An asymptotic analysis for a large class of stochastic optimization problems arising in manufacturing is presented. A typical example of the problems considered in this paper is a production planning problem with random capacity and demand. In this example, it is assumed that the capacity of the system fluctuates faster than the other quantities. The general model considered here also has a fast controlled Markov process in its state description. By using the difference in the time scales of different quantities, the problem is simplified by “averaging” out the fast process. Then asymptotically optimal strategies are constructed from the optimal solutions of the limiting problems. The proofs of these results use the theory of viscosity solutions to dynamic programming equations. However, the formal construction of the asymptotically optimal strategies does not require knowledge of this theory.

Key words. dynamic programming, viscosity solutions, production planning, manufacturing, singular perturbations

AMS(MOS) subject classifications. 90B30, 93E20, 35R35

1. Introduction. Most modern manufacturing systems are complex and large in scale, including several subsystems, a wide variety of equipment, and a number of different products. Moreover, operating policies of these systems must respond to discrete events that are quite different from one another, for example, machine setups, failure and repairs, demand changes, purchasing and building new facilities, etc. Because of the size of the systems, it is impossible to achieve optimal operating policies. The only practical strategies are the suboptimal ones, derived using the structure of a given system. Generally, these techniques amount to reduction of the complexity by decomposing the original system into simpler subsystems. We limit ourselves to systems that have hierarchical decomposition. Based on this structure we “average out” certain parameters, thus simplifying the optimization problems. Then suboptimal policies are obtained as solutions to these simplified problems. For further information on control of manufacturing systems, we refer the reader to Gershwin et al. [9]; on hierarchical production planning, see Gershwin [8] and Bitran and Tirupati [3].

Recently Lehoczy et al. [11] carried out the above procedure for a specific stochastic production planning problem. However, the scope of the mathematical tools used in [11] is not limited to the production planning problem. In this paper, we demonstrate the versatility of these techniques by introducing a general framework for asymptotic analysis of optimal stochastic control problems. This framework, in particular, includes the problem studied in [11] and its generalizations.

Typical of the problems we consider is a production planning problem subject to random changes in capacity and demand. We consider the case in which the capacity fluctuates faster than the other quantities, when the system is working. In other words, when there is production, the rate at which the capacity changes occur is much larger than the rate of fluctuation in demand, the rate of discounting, and other time scales. In this model the capacity process depends on the production rate. The model without this dependency is analyzed in [11] and a limiting problem is obtained by simply replacing the random capacity by its average value. However, for the general model,

* Received by the editors December 5, 1990; accepted for publication (in revised form) September 23, 1991. This research was supported in part by National Science Foundation grant DMS-9002249 and Army Research Office grant DAAL-03-86-K-0171 and the Center for Nonlinear Analysis.

† Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

a straightforward averaging, as it was done in [11], is no longer valid. In fact, the “average” capacity is a function of the production rate and its computation is quite complicated. This dependency also implies that in general the diffusion approximation is not possible. Therefore, we are not able to use the elegant analysis of Kushner [10].

The mathematical analysis of this paper uses the dynamic programming principle and the viscosity solutions of the differential equations. Although our proofs are complicated at times, on the formal level the methodology is straightforward and we wish to emphasize this. An outline of the formal method is as follows: First derive the dynamic programming (Bellman) equation for the full problem. Then let the fluctuation rate of the faster process go to infinity in the equation. Obtain the formal limiting equation by assuming the regularity of the value function. Compute the optimal control problem related to the limiting equation and its optimal solution. This solution in turn generates an asymptotically optimal control for the original model. The asymptotic optimality of this control was recently proved by Zhang and Sethi [15] for the model considered in [11]. Finally, we note that our techniques are related to those in Bensoussan [2] and Saksena, O’Reilly, and Kokotovic [13].

The paper is organized as follows. The stochastic production planning problem is described in § 2. Using this problem as a model, we introduce the general framework in § 3. Section 4 is devoted to the proof of the convergence result. A suboptimal but asymptotically optimal control is constructed in § 5. Finally, a discussion of the convergence rate is given.

2. Production planning. Consider a manufacturing facility in which there are m identical machines that are equally capable of producing n distinct part types. The production must be scheduled to meet a demand that fluctuates randomly. However, we assume that the machines are subject to a Markovian breakdown and repair process. Thus the demand may not be met every time, and the production strategies should take this into account.

Akella and Kumar [1] studied the one-dimensional model ($n = m = 1$) with a constant demand rate. They explicitly computed the optimal production rate, which is a bang-bang control. They showed that there is a threshold level $\alpha^* \geq 0$ such that, when the only available machine is in working condition, the production rate is either zero or equal to the full capacity if the inventory is strictly greater than α^* or less than α^* . Of course, when the machine is down, the only possible production rate is zero.

The general model we are considering also admits an optimal control, which is bang-bang. However, for large n and m the computation of the threshold levels is complicated. Also the description of the production rate includes not only the threshold levels but the fractions of the capacity devoted to each part. We simplify this model by using its hierarchical structure. As discussed in the Introduction, we assume that the occurrence of machine breakdown and repair process is faster than the other time scales that are relevant to this problem.

We continue with the description of the model. Let an n -vector $x(t)$ denote the inventory at time $t \geq 0$. For a given production rate (control) $u(t)$, the inventory (state) satisfies the ordinary differential equation

$$(2.1) \quad \frac{d}{dt} x(t) = u(t) - d(t), \quad t > 0,$$

where $d(t) = (d_1(t), \dots, d_n(t))$ is the demand vector. The demand process is assumed to be Markov, taking values in a discrete set $D \subset (0, \infty)^n$. The components of the production rate are nonnegative and they are bounded from above by a constant related

to $\alpha^\varepsilon(t)$, the number of available machines at time t . We assume that $\alpha^\varepsilon(t) \in \{0, 1, \dots, m\}$ is a Markov chain with infinitesimal generator

$$(2.2) \quad \frac{1}{\varepsilon} Q^{u(t)} = \left(\frac{1}{\varepsilon} q_{ij}(u(t)) \right)_{i,j=0,1,\dots,m}.$$

Note that the generator of α^ε depends on the production rate $u(t)$. Since the machine failures are more likely when the production rate is high, this dependence is a natural one. However, in certain situations one may argue that it is negligible as it was assumed in [11].

The parameter $\varepsilon > 0$ appearing in the machine availability process is related to the hierarchy in the time scales. Indeed, the mean rate of change of $\alpha^\varepsilon(\cdot)$ is of order $1/\varepsilon$, while the rate of change of demand is bounded in ε . Hence, for small $\varepsilon > 0$, these two time scales are of different order.

The optimization problem is to *minimize*

$$(2.3) \quad J^\varepsilon(x, d, i; u) = E_{x,d,i} \int_0^\infty e^{-t} G(x(t), u(t)) dt$$

over all nonanticipative production processes, $u(t)$, satisfying the machine availability constraint

$$(2.4) \quad u(t) \in K(\alpha^\varepsilon(t)) \quad \forall t \geq 0,$$

where $E_{x,d,i}$ denotes the mathematical expectation with initial conditions $x(0) = x$, $d(0) = d$, and $\alpha^\varepsilon(0) = i$. The constraint set is given by

$$K(i) = \left\{ u \in [0, \infty)^n : \sum_{k=1}^n u_k \gamma_k \leq i \right\}, \quad i = 1, \dots, m,$$

with nonnegative constants γ_k .

Let $v^\varepsilon(x, d, i)$ be the value function

$$v^\varepsilon(x, d, i) = \inf_{u(\cdot)} J^\varepsilon(x, d, i; u), \quad x \in R^n, \quad d \in D, \quad i \in \{0, 1, \dots, m\}.$$

Then v^ε is a (viscosity) solution of

$$(2.5) \quad \begin{aligned} 0 = v^\varepsilon(x, d, i) + \sup_{u \in K(i)} & \left\{ -(u-d) \cdot D_x v^\varepsilon(x, d, i) - G(x, u) \right. \\ & - \frac{1}{\varepsilon} \sum_{j=0}^m q_{ij}(u) [v^\varepsilon(x, d, j) - v^\varepsilon(x, d, i)] \\ & \left. - \sum_{d' \in D} \bar{q}_{dd'} [v^\varepsilon(x, d', i) - v^\varepsilon(x, d, i)] \right\} \end{aligned}$$

for all $x \in R^n$, $d \in D$, $i \in \{0, 1, \dots, m\}$, where $\bar{Q} = (\bar{q}_{dd'})_{d,d' \in D}$ is the infinitesimal generator of $d(\cdot)$, and D_x denotes the gradient in the x -variable.

We close this section by rewriting (2.5) in a manner which is compatible with the notation of the next section. For $(x, d, i) \in R^n \times D \times \{0, 1, \dots, m\}$, and $p \in R^n$, $L \in R^{|D|}$, $\kappa \in R^{m+1}$, define $H(x, d, i; p, L, \kappa)$ by

$$(2.6) \quad \begin{aligned} H(x, d, i; p, L, \kappa) = \sup_{v \in K(i)} & \left\{ -(u-d) \cdot p - G(x, u) - \sum_{j=0}^m q_{ij}(u) [\kappa_j - \kappa_i] \right\} \\ & - \sum_{d' \in D} \bar{q}_{dd'} [L_{d'} - L_d]. \end{aligned}$$

Then (2.5) can be rewritten as

$$v^\varepsilon(x, d, i) + H\left(x, d, i; D_x v^\varepsilon(x, d, i), v^\varepsilon(x, \cdot, i), \frac{1}{\varepsilon} v^\varepsilon(x, d, \cdot)\right) = 0.$$

Finally, we note that the sum of the entries of each row of any infinitesimal generator is zero, i.e.,

$$\sum_{j=0}^m q_{ij}(u) = \sum_{d' \in D} \bar{q}_{dd'} = 0, \quad \forall i \in \{0, \dots, m\}, \quad d \in D.$$

This implies that

$$(2.7) \quad H(x, d, i; p, L + c_1, \kappa + c_2) = H(x, d, i; p, L, \kappa)$$

for any constants $c_1, c_2 \in (-\infty, \infty)$, and $L + c_1$ denotes the vector obtained by adding the constant c_1 to each component of L ; $\kappa + c_2$ is defined similarly.

3. General model. We consider a family of discounted, infinite horizon stochastic optimal control problems indexed by a parameter $\varepsilon > 0$, with a state space $\Sigma = R^n \times D \times Z$. We take both D and Z to be finite sets. For $(x, d, i) \in \Sigma$, let $v^\varepsilon(x, d, i)$ be the value function satisfying the dynamic programming equation

$$(3.1) \quad v^\varepsilon(x, d, i) + H\left(x, d, i; D_x v^\varepsilon(x, d, i), v^\varepsilon(x, \cdot, i), \frac{1}{\varepsilon} v^\varepsilon(x, d, \cdot)\right) = 0 \quad \forall (x, d, i) \in \Sigma,$$

where H is a real valued function of $\Sigma \times R^n \times R^{|D|} \times R^{|Z|}$. We will not describe the underlying stochastic model. But the function H is given in terms of the running cost and the dynamics of the state process. In particular H is *jointly convex in the last three variables and has the invariance property* (2.7). We now make a structural assumption. Fix $(x, d) \in R^n \times D$, $p \in R^n$, $L \in R^{|D|}$, and $\alpha \in R^{|Z|}$. Consider the nonlinear equation

$$(3.2) \quad \alpha_i + H(x, d, i; p, L, \kappa) = 0 \quad \forall i \in Z,$$

where $\kappa \in R^{|Z|}$ is the unknown. Due to the translation invariance (2.7), if κ is a solution of (3.2) then $\kappa + c$ is a solution for any constant c . So we should search for a unique solution in the quotient space which we call

$$(3.3) \quad \mathcal{P}_{|Z|} = \left\{ \kappa \in R^{|Z|} : \sum_{i \in Z} \kappa_i = 0 \right\}.$$

The translation invariance also yields that the range of the map $\kappa \mapsto \{H(x, d, i; p, L, \kappa)\}_{i \in Z}$ is not equal to $R^{|Z|}$. Hence we may only expect (3.2) to have a unique solution $\kappa \in \mathcal{P}_{|Z|}$ provided that the components of α satisfy a (possibly nonlinear) scalar equation. More precisely we assume that there are functions

$$(3.4i) \quad H_{av} : R^n \times D \times R^n \times R^{|D|} \times R^{|Z|} \rightarrow R,$$

and

$$(3.4ii) \quad A : \Sigma \times R^n \times R^{|D|} \times R^{|Z|} \rightarrow R$$

such that for all $(x, d, i) \in \Sigma$, $p \in R^n$, $L \in R^{|D|}$, and $\alpha \in R^{|Z|}$, we have $A(x, d, \cdot; p, L; \alpha) \in \mathcal{P}_{|Z|}$, and

$$(3.5) \quad \alpha_i + H(x, d, i; p, L, A(x, d, \cdot; p, L; \alpha)) = 0,$$

provided that $\alpha = \{\alpha_i\}_{i \in Z}$ satisfies

$$(3.6) \quad H_{av}(x, d; p, L; \alpha) = 0.$$

Clearly, the function H_{av} is not uniquely determined. However, under mild assumptions on the coefficients of the optimization problem, we can show that it is continuous and monotone in α . Then by multiplying it by (-1) if necessary, we may take it to be nondecreasing in α . So the following assumption is not restrictive:

(3.7) A, H_{av} are continuous and H_{av} is nondecreasing in α .

Note that (3.1) is similar to (3.2). However, in (3.2) variables p and L are assumed to be independent of i , and in (3.1) $p = D_x v^\varepsilon(x, d, i)$ and $L_{d'} = v^\varepsilon(x, d', i)$. However, we expect the dependence of v^ε on i to be averaged out in the limit $\varepsilon \rightarrow 0$. So suppose that $v^\varepsilon(x, d, i)$ converges to $v(x, d)$, and

$$\kappa^\varepsilon(x, d, j) = \frac{1}{\varepsilon} \left[v^\varepsilon(x, d, j) - \sum_{k \in Z} v^\varepsilon(x, d, k) \right]$$

converges to $\kappa(x, d, j)$. Due to the invariance (2.7), we may rewrite (3.1) as

$$v^\varepsilon(x, d, i) + H(x, d, i; D_x v^\varepsilon(x, d, i), v^\varepsilon(x, \cdot, i), \kappa^\varepsilon(x, d, \cdot)) = 0.$$

Now let ε go to zero. Formally, we obtain

$$v(x, d) + H(x, d, i; D_x v(x, d), v(x, \cdot), \kappa(x, d, \cdot)) = 0 \quad \forall i \in Z.$$

Note that the above equation is a special case of (3.2) with $p = D_x v(x, d)$ and $L_{d'} = v(x, d')$. Hence (3.6) yields

$$\bar{H}_{av}(x, d; D_x v(x, d), v(x, \cdot); v(x, d)) = 0,$$

where for $x \in R^n$, $d \in D$, $p \in R^n$, $L \in R^{|D|}$, and a scalar v ,

$$\bar{H}_{av}(x, d; p, L; v) = H_{av}(xd; p, L; \bar{v})$$

with $\bar{v} = (v, \dots, v) \in R^{|Z|}$.

In the next section, we will show that v^ε converges to a solution of the above equation. Since \bar{H}_{av} is convex in the last three variables, $\bar{H}_{av} = 0$ is the dynamic programming equation of an optimal control problem with state space $R^n \times D$. Therefore v is the value function of this problem. The connection between the equation $\bar{H}_{av} = 0$ and the optimal control problem will be clarified in Examples 3.1–3.3, below.

Our final assumption is a strong monotonicity condition on H_{av} . For each $i \in Z$, $\kappa^i \in R^{|Z|}$, set

$$\alpha_i = -H(x, d, i; p, L, \kappa^i), \quad i \in Z.$$

Since κ^i may depend on i , we can *not* conclude that (3.6) holds. However, we assume that

$$(3.8i) \quad H_{av}(x, d, i; p, L; \alpha) \leq 0 \quad (\text{or } \geq 0, \text{ respectively})$$

whenever there is $\kappa \in R^{|Z|}$ such that for all $i, j \in Z$,

$$(3.8ii) \quad \kappa_j^i \leq \kappa_j^i + [\kappa_i - \kappa_j] \quad (\text{or } \geq, \text{ respectively}).$$

We now give two examples to clarify the above hypothesis.

Example 3.1. Consider (2.5) with $n = m = 1$, $D = \{d_0\}$, and $q_{01}(u) = -q_{00} = \lambda > 0$, $q_{10}(u) = -q_{11}(u) = \mu(u) \geq 0$. Then $K(0) = \{0\}$, $K(1) = [0, 1/\gamma_1]$ and the Hamiltonian H

in (2.6) has the form

$$H(x, 0; p, \kappa) = d_0 p - G(x, 0) - \lambda[\kappa_1 - \kappa_0],$$

$$H(x, 1; p, \kappa) = \sup_{0 \leq u \leq 1/\gamma_1} \{-up - G(x, u) + \mu(u)[\kappa_1 - \kappa_0]\} + d_0 p.$$

Equation (3.2) is equivalent to

$$(3.9i) \quad \alpha_0 + d_0 p - G(x, 0) - \lambda[\kappa_1 - \kappa_0] = 0,$$

$$(3.9ii) \quad \alpha_1 + \sup_{0 \leq u \leq 1/\gamma_1} \{-(u - d_0)p - G(x, u) + \mu(u)[\kappa_1 - \kappa_0]\} = 0.$$

Suppose that for a given (α_0, α_1) we have a solution $(\kappa_0, \kappa_1) \in \mathcal{P}_2$ solves (3.9). Then (3.9i) yields

$$(3.10) \quad \kappa_1 - \kappa_0 = \frac{1}{\lambda} [\alpha_0 + d_0 p - G(x, 0)].$$

Since $(\kappa_0, \kappa_1) \in \mathcal{P}_2$, $\kappa_0 + \kappa_1 = 0$. Therefore,

$$\kappa_1 = A(x, 1; p, \alpha) = \frac{1}{2\lambda} [\alpha_0 + d_0 p - G(x, 0)],$$

$$\kappa_0 = A(x, 0; p, \alpha) = -\kappa_1.$$

Observe that we used only (3.9i) to obtain the above formula. The other equation, (3.9ii), will be used to compute H_{av} . Indeed, using (3.10) in (3.9ii), we arrive at

$$H_{av}(x; p; \alpha) = 0,$$

where

$$H_{av}(x; p; \alpha_0, \alpha_1) = \alpha_1 + \sup_{0 \leq u \leq 1/\gamma_1} \left\{ -(u - d_0)p - G(x, u) + \frac{\mu(u)}{2\lambda} [\alpha_0 + d_0 p - G(x, 0)] \right\}.$$

To verify (3.8), suppose that (3.9i) holds with $\kappa^0 = (\kappa_0^0, \kappa_1^0)$ and (3.9ii) holds with $\kappa^1 = (\kappa_0^1, \kappa_1^1)$. Then (3.10) holds with $\kappa_1^0 - \kappa_0^0$ on the left-hand side. Also suppose that (3.8ii) holds. Then

$$\kappa_1^0 - \kappa_0^0 \leq \kappa_1^1 - \kappa_0^1.$$

Using the above inequality and (3.10) in (3.9ii), we obtain $H_{av}(x, p; \alpha_0, \alpha_1) \leq 0$. Hence (3.8i) holds.

In this example, the optimal control problem related to the limiting equation $\bar{H}_{av} = 0$ is to minimize

$$\int_0^\infty \exp\left(-t - \int_0^t \frac{\mu(u(s))}{2\lambda} ds\right) \left[G(x(t), u(t)) + \frac{\mu(u(t))}{2\lambda} G(x(t), 0) \right] dt$$

subject to

$$\frac{d}{dt} x(t) = u(t) - \left(1 + \frac{\mu(u(t))}{2\lambda}\right) d_0, \quad t > 0$$

and $u(t) \in [0, 1/\gamma_1]$, $t \geq 0$.

Example 3.2. Again consider (2.5) with $n = 1$, $m = 2$, $D = \{d_0\}$, and

$$Q(u) = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 \\ \mu_1(u) & -[\mu_1(u) + \lambda_2] & \lambda_2 \\ 0 & \mu_2(u) & -\mu_2(u) \end{bmatrix}.$$

Then $K(0) = \{0\}$, $K(1) = [0, 1/\gamma_1]$, $K(2) = [0, 2/\gamma_1]$. Suppose that for a given $x, p, \alpha \in \mathbb{R}^3$, $K \in \mathcal{P}_3$ solves (8.2). Then a computation similar to the previous case yields

$$\begin{aligned}\kappa_1 - \kappa_0 &= \frac{1}{\lambda_1} [\alpha_0 + d_0 p - G(x, 0)], \\ \kappa_2 - \kappa_1 &= \frac{1}{\lambda_2} \left[\alpha_1 + \sup_{0 \leq u_1 \leq 1/\gamma_1} \{-(u_1 - d_0)p - G(x, u_1) + \mu_1(u_1)[\kappa_1 - \kappa_0]\} \right], \\ \alpha_2 + \sup_{0 \leq u_2 \leq 2/\gamma_1} \{-(u_2 - d_0)p - G(x, u_2) + \mu_2(u_2)[\kappa_2 - \kappa_1]\} &= 0.\end{aligned}$$

Hence

$$H_{av}(x; p; \alpha) = \sup_{\substack{0 \leq u_1 \leq 1/\gamma_1 \\ 0 \leq u_2 \leq 2/\gamma_1}} \left\{ -f(u_1, u_2)p - g(x, u_1, u_2) + \alpha_2 + \frac{\mu_2(u_2)}{\lambda_2} \left[\alpha_1 + \alpha_0 \frac{\mu_1(u_1)}{\lambda_1} \right] \right\},$$

where

$$\begin{aligned}f(u_1, u_2) &= (u_2 - d_0) + \frac{\mu_2(u_2)}{\lambda_2} \left[(u_1 - d_0) - \frac{\mu_1(u_1)}{\lambda_1} d_0 \right], \\ g(x, u_1, u_2) &= G(x, u_2) + \frac{\mu_2(u_2)}{\lambda_2} \left[G(x, u_1) + \frac{\mu_1(u_1)}{\lambda_1} G(x, 0) \right].\end{aligned}$$

The corresponding control problem is similar to that described in the Example 3.1.

Example 3.3. Again consider (2.5) with $n = 1$, $D = \{d_0\}$ and $Q(u) = Q$ is an irreducible $(m+1) \times (m+1)$ stochastic matrix. Then, (3.2) has the form

$$(3.11) \quad \alpha_i = - \sup_{0 \leq u \leq i/\gamma_i} \{-(u - d_0) \cdot p - G(x, u)\} + (Q\kappa)_i, \quad i \in \{0, 1, \dots, m\}.$$

Since Q is irreducible, there is a positive vector $\nu \in \mathbb{R}^{m+1}$ such that $\nu_i > 0$, $\sum_i \nu_i = 1$, and $(\nu Q)_i = 0$ for all i . Multiply the above equation by ν_i and sum over i to obtain

$$\sum_{i=0}^m \left[\alpha_i \nu_i + \sup_{0 \leq u_i \leq i/\gamma_i} \{ -\nu_i(u_i - d_0)p - \nu_i G(x, u_i) \} \right] = 0.$$

A straightforward algebraic manipulation gives

$$H_{av}(x, p; \alpha) = \sum_{i=0}^m \alpha_i \nu_i + \sup_{0 \leq u \leq \bar{v}} \{ -(u - d_0)p - \bar{G}(x, u) \},$$

where

$$\begin{aligned}\bar{v} &= \sum_{i=0}^m i \nu_i / \gamma_i, \\ \bar{G}(x, u) &= \inf \left\{ \sum_{i=0}^m \nu_i G(x, u_i) : u_j \in K(j) \text{ and } \sum_{i=0}^m \nu_i u_i = u \right\}.\end{aligned}$$

To verify (3.8), suppose that (3.11) holds with $\kappa^i \in \mathbb{R}^{m+1}$, and κ^i 's satisfy (3.8ii). Multiply (3.11) by ν_i , sum over i , and then use the formula for H_{av} to obtain

$$H_{av}(x, p; \alpha) = \sum_{i,j=0}^m \nu_i q_{ij} \kappa_j^i.$$

Now use (3.8ii) and the nonnegativity of q_{ij} for $i \neq j$, to obtain

$$\begin{aligned} \sum_{i,j=0}^m \nu_i q_{ij} \kappa_j^i &\leq \sum_{i,j=0}^m \nu_i q_{ij} [\kappa_j^j + (\kappa_j - \kappa_i)] \\ &= \sum_{j=0}^m (\kappa_j^j + \kappa_j) \left[\sum_{i=0}^m \nu_i q_{ij} \right] - \sum_{i=0}^m (\nu_i \kappa_i) \left[\sum_{j=0}^m q_{ij} \right] \\ &= 0. \end{aligned}$$

The corresponding optimal control problem is simple, and it is described in Example 5.1.

These examples can easily be generalized to obtain the following lemma.

LEMMA 3.1. *Suppose that H is as in (2.4) and $G(x, u)$ is convex in the u -variable, and either $\alpha^\varepsilon(t)$ is a birth-death process, i.e.,*

$$(3.12) \quad q_{ij}(u) = \begin{cases} \mu_i(u), & j = i - 1, \quad i = 1, \dots, m, \\ \lambda_i, & i = j - 1, \quad j = 1, \dots, m, \\ -[\lambda_i + \mu_i(u)], & \text{if } i = j = 0, \dots, m, \\ 0, & \text{otherwise,} \end{cases}$$

with $\mu_i(u) \geq 0$, $\lambda_i > 0$, or $Q(u) \equiv Q$ for all u and Q is irreducible. Then the assumptions (3.5), (3.6), and (3.8) are satisfied.

The convergence results under the second hypothesis is first obtained in [11]. These results are then improved in [15]. The asymptotic analysis of v^ε under the first set of assumptions, however, is not covered in the previous studies. In this case, the parameter λ_i is the machine repair rate when $i - 1$ machine are operating. It is natural to assume that λ_i is independent of the production rate. The quantity $\mu_i(u)$ is the machine failure rate when i machines are operating with a production rate of u , and, in general, μ_i is a function of the production rate.

4. Convergence. In this section we study the limiting behavior of v^ε as ε tends to zero. In whatever follows we always assume the structural assumptions (3.5)–(3.8). However, to obtain convergence results we need to impose some uniform estimates on v^ε . In this section we assume that there are $K, \nu \geq 0$, independent of ε , such that for all $\varepsilon \in (0, 1]$, $(x, d, i) \in \Sigma$,

$$(4.1i) \quad |v^\varepsilon(x, d, i)| \leq K(1 + |x|^\nu),$$

$$(4.1ii) \quad \frac{1}{|x - y|} |v^\varepsilon(x, d, i) - v^\varepsilon(y, d, i)| \leq K(1 + |x|^\nu), \quad 0 < |y - x| \leq 1,$$

$$(4.1iii) \quad |v^\varepsilon(x, d, i) - v^\varepsilon(x, d, j)| \leq \varepsilon K(1 + |x|^\nu), \quad j \in \mathbb{Z}.$$

The inequality (4.1ii) is a uniform Lipschitz estimate. If the function v^ε is continuously differentiable in the x -variable, then (4.1ii) is equivalent to the uniform boundedness of the gradient, i.e.,

$$(4.2) \quad \sup_{\varepsilon \in (0, 1]} |D_x v^\varepsilon(x, d, i)| \leq K(1 + |x|^\nu).$$

The estimate (4.1iii) is related to the scaling used in the equation (3.1). Notice that in (3.1) the vector $(1/\varepsilon)v^\varepsilon(x, d, \cdot)$ appears. So intuitively we expect the differences $|v^\varepsilon(x, d, i) - v^\varepsilon(x, d, j)|/\varepsilon$ to be locally bounded as assumed in (4.1iii). Note that the translation invariance (2.7) is the reason why we do not expect $v^\varepsilon/\varepsilon$ to be bounded.

In the production planning examples, these estimates are always satisfied. Indeed consider the cases discussed in Lemma 3.1 and assume that

$$(4.3) \quad |G(x, u)| + \frac{1}{|x - y|} |G(x, u) - G(y, u)| \leq K(1 + |x|^\nu)$$

for all $x, y, u \in R^n$, $0 < |x - y| \leq 1$. Then for the second case of Lemma 3.1, the estimates (4.1) are proved in [11]; see Lemma 2.1 in [11]. A very similar proof yields these estimates also in the first case of Lemma 3.1.

Using (4.1) and the Ascoli–Arzela theorem we construct a sequence, denoted by ε again, and locally Lipschitz continuous function $v(x, d)$ such that $v^\varepsilon(x, d, i)$ converges to $v(x, d)$ uniformly on compact subsets on Σ . As we discussed in § 3, formally v solves the limiting equation

$$(4.4) \quad \bar{H}_{av}(x, d; D_x v(x, d), v(x, \cdot); v(x, d)) = 0, \quad (x, d) \in R^n \times D.$$

Recall that for $(x, d; p, L) \in R^n \times D \times R^n \times R^{|D|}$ and a scalar v ,

$$\bar{H}_{av}(x, d; p, L; v) := H_{av}(x, d; p, L; \bar{v}),$$

with $\bar{v} = (v, v, \dots, v) \in R^{|Z|}$. We will show below that v indeed is a solution of (4.4).

In general, v is not differentiable and the equation (4.4) must be interpreted in the viscosity sense. We refer the reader to Crandall and Lions [5]; Crandall, Evans, and Lions [4]; Lions [12]; Soner [14]; Fleming, Sethi, and Soner [6]; and [11] for the definition and the properties of the viscosity solutions of (3.1) or (4.4).

THEOREM 4.1 (Stability). *Assume (4.1), and that v^ε is a viscosity solution of (3.1). Suppose that (4.4) has a unique viscosity solution v satisfying (4.1). Then v^ε converges to v uniformly on compact subsets of $R^n \times D$, as ε tends to zero.*

Proof. Let $\bar{v}(x, d)$ be the limit of $v^{\varepsilon_m}(x, d, i)$ for some sequence $\varepsilon_m \rightarrow 0$. Let $\psi(x, d)$ be a continuously differentiable function and for $d \in D$, let $x_0 \in R^n$ be the strict maximum of $\bar{v}(\cdot, d) - \psi(\cdot, d)$ on R^n . To show that \bar{v} is a viscosity subsolution of (4.4), we must verify the inequality

$$(4.5) \quad \bar{H}_{av}(x_0, d; D_x \psi(x_0, d), \bar{v}(x_0, \cdot); \bar{v}(x_0, d)) \leq 0.$$

Consider the map $x \mapsto v^{\varepsilon_m}(x, d, i) - \psi(x, d)$. Since x_0 is a strict maximizer, there are $x_m(i) \in R^n$ converging to x_0 and maximizing the above map locally in the x variable. Then the viscosity property of $v^m := v^{\varepsilon_m}$ yields

$$v^m(x_m(i), d, i) + H\left(x_m(i), d, i; D_x \psi(x_m(i), d), v^m(x_m(i), \cdot, i), \frac{1}{\varepsilon_m} v^m(x_m(i), d, \cdot)\right) \leq 0.$$

Since v^m converges to \bar{v} and $x_m(i)$ converges to x_0 , there is a sequence $K_m \rightarrow 0$ such that

$$(4.6) \quad \bar{v}(x_0, d) + H(x_0, d, i; p_0, \bar{v}(x_0, \cdot), \kappa^{m,i}) \leq K_m,$$

where $p_0 = D_x \psi(x_0, d)$ and

$$\kappa_j^{m,i} = \frac{1}{\varepsilon_m} v^m(x_m(i), d, j), \quad x \in R^n, \quad i, j \in Z.$$

Since $x_m(i)$ is a local maximizer of $v^m(\cdot, d, i) - \psi(\cdot, d)$, we have

$$v^m(x_m(j), d, j) - \psi(x_m(j), d) \geq v^m(x_m(i), d, j) - \psi(x_m(i), d)$$

for all $i, j \in Z$. Set $\kappa_j = \psi(x_m(j), d)$. Then

$$\kappa_j^{m,i} \leq \kappa_j^{m,j} + (\kappa_i - \kappa_j)$$

for every $i, j \in Z$. Hence, (3.8) implies that

$$H_{av}(x_0, d; p_0, \bar{v}(x_0, \cdot); \beta) \leq 0,$$

where

$$\beta(i) = -H(x_0, d, i; p_0, \bar{v}(x_0, \cdot), \kappa^{m,i}), \quad i \in Z.$$

Also (4.6) yields that $\beta(i) \geq \bar{v}(x_0, d) - K_m$ for every $i \in Z$. Hence, the monotonicity of H_{av} yields that

$$\bar{H}_{av}(x_0, d; p_0, \bar{v}(x_0, \cdot); \bar{v}(x_0, d) - K_m) \leq H_{av}(x_0, d; p_0, \bar{v}(x_0, \cdot); \beta) \leq 0 \quad \forall m.$$

Now let m go to infinity and use the fact that $K_m \rightarrow 0$ to obtain (4.5). Hence \bar{v} is a viscosity subsolution. Similarly we can show that it is also a viscosity supersolution, and therefore a solution. Since (4.4) has a unique viscosity solution v satisfying (4.1), $\bar{v} = v$. \square

COROLLARY 4.1. *Assume the hypothesis of Lemma 3.1, and (4.3). Then v^ε converges uniformly on compact subsets of Σ , as ε tends to zero.*

Proof. We have argued that (4.3) implies the estimates (4.1). Also the uniqueness of viscosity solutions of (4.4) satisfying (4.1) follows from the classical techniques of Crandall, Evans, Lions [4]. \square

5. Asymptotically optimal controls. In this section we outline a procedure of constructing suboptimal controls by using the limiting equation (4.4). We will show that under certain assumptions, the difference between the value function and the performance of the controls that we construct converges to zero in the limit $\varepsilon \rightarrow 0$. Before we describe the procedure for the general case, we discuss two examples.

Example 5.1. Consider the case described in Example 3.3. Let $v(x)$ be the unique viscosity solution of the limit equation,

$$\begin{aligned} 0 &= \bar{H}_{av}(x, D_x v(x), v(x)) \\ &= v(x) + \sup_{0 \leq u \leq \bar{v}} \{-(u - d_0)D_x v(x) - \bar{G}(x, u)\}. \end{aligned}$$

Then $v(x)$ is the value function of a deterministic optimal control problem. Indeed,

$$v(x) = \inf \int_0^\infty e^{-t} \bar{G}(x(t), u(t)) dt,$$

subject to constraints $x(0) = x$, (2.1) with $d(t) \equiv d_0$, and $0 \leq u(t) \leq \bar{v}$ for all $t \geq 0$. Suppose that v is differentiable. For each x , pick

$$u^*(x) \in \operatorname{argmax} \{-(u - d_0)D_x v(x) - \bar{G}(x, u) : 0 \leq u \leq \bar{v}\}.$$

If

$$(5.1) \quad \frac{d}{dt} x(t) = u^*(x(t)) - d_0, \quad t > 0,$$

has a solution, then it is elementary to show that $\tilde{u}(t) = u^*(x(t))$ is optimal. So suppose that this is the case. Then we construct a feedback control for the $\varepsilon > 0$ problem by setting

$$u^{*,\varepsilon}(x, i) = iu^*(x), \quad x \in (-\infty, \infty), \quad i = 0, 1, \dots, m.$$

Then we expect $u^{*,\varepsilon}$ to perform close to the optimal control. Indeed, Zhang and Sethi [15] have shown that

$$\lim_{\varepsilon \rightarrow 0} |J^\varepsilon(x, i; u^{*,\varepsilon}) - v^\varepsilon(x, i)| = 0,$$

provided that (5.1) has a unique solution.

Example 5.2. Now we return to Example 3.2. As in the previous example suppose that the solution v of the limit equation is differentiable, i.e.,

$$(5.2) \quad \sup_{\substack{0 \leq u_1 \leq 1/\gamma_1 \\ 0 \leq u_2 \leq 2/\gamma_2}} \{l(u_1, u_2)v(x) - f(u_1, u_2)D_x v(x) - g(x, u_1, u_2)\} = 0,$$

where f, g are as in Example 3.2, and

$$l(u_1, u_2) = 1 + \frac{\mu_2(u_2)}{\lambda_2} \left[1 + \frac{\mu_1(u_2)}{\lambda_1} \right].$$

Let $u_1^*(x), u_2^*(x)$ be a maximizer in (5.2). Clearly, the sequence $u^*(x, 0) = 0, u^*(x, 1) = u_1^*(x)$, and $u^*(x, 2) = u_2^*(x)$ satisfies the machine availability constraint and is a candidate for an asymptotically optimal control. We will show in Theorem 5.1 below that this is indeed the case, provided that u^* has certain properties.

To motivate the construction in the general framework, we will derive a property of u^* next. Set

$$p(x) = D_x v(x),$$

and

$$A(x, i) = A(x, i; p(x), \bar{v}(x)), \quad i = 0, 1, 2;$$

recall that $\bar{v}(x) = (v(x), v(x), v(x))$. Then by (3.5), we have

$$\begin{aligned} 0 &= v(x) + H(x, i; p(x), A(x, \cdot)) \\ &= v(x) + \sup_{0 \leq u \leq i/\gamma_1} \{- (u - d_0)p(x) - G(x, u) - (Q(u)A(x, \cdot))(i)\}, \quad i = 0, 1, 2. \end{aligned}$$

Then it is straightforward to show that $u^*(x, i)$ maximizes the expression in the above equation. We will use this description of u^* in the discussion of the general problem.

In general, the Hamiltonian H has the form

$$(5.3) \quad H(\xi; p, L, \kappa) = \sup_{u \in K(\xi)} \{-\mathcal{L}^{u,\xi}(p, L, \kappa) - G(\xi, u)\}$$

for $\xi \in \Sigma, p \in \mathbb{R}^n, L \in \mathbb{R}^{|D|}, \kappa \in \mathbb{R}^{|Z|}$, a set $K(\xi) \subset U$, a function G of $\Sigma \times U$, and a family of linear operators $\mathcal{L}^{u,\xi}(p, L, \kappa)$, which are invariant under scalar translations of L, κ . In the notation of § 2, for example, $U = [0, \infty)^n, K(\xi) = K(i)$,

$$(5.4) \quad \mathcal{L}^{u,\xi}(p, L, \kappa) = (u - d) \cdot p + (\bar{Q}L)(d) + (Q(u)\kappa)(i)$$

for $\xi = (x, d, i) \in \Sigma$. Assume that v is differentiable and set

$$p(x, d) = D_x v(x, d), \quad A(x, d, i) = A(x, d, i; p(x, d), v(x, \cdot); \bar{v}(x, d)).$$

Then choose $u^*(\xi) \in K(\xi)$ such that

$$\begin{aligned} (5.5) \quad & -\mathcal{L}^{u^*(\xi),\xi}(p(x, d), v(x, \cdot), A(x, d, \cdot)) - G(\xi, u^*(\xi)) \\ & = H(\xi; p(x, d), v(x, \cdot), A(x, d, \cdot)). \end{aligned}$$

It is known that not every feedback control yields a well-defined state process. However, we assume that u^* is indeed related to a well-defined state process. Let $J^{\varepsilon,*}(\xi)$ be the value of the pay-off functional. Then $J^{\varepsilon,*}$ formally solves

$$(5.6)^{\varepsilon} \quad J^{\varepsilon,*}(\xi) - \mathcal{L}^{u^*(\xi),\xi} \left(D_x J^{\varepsilon,*}(\xi), J^{\varepsilon,*}(x, \cdot, i), \frac{1}{\varepsilon} J^{\varepsilon,*}(x, d, \cdot) \right) - G(\xi, u^*(\xi)) = 0.$$

The definition of u^* implies that the formal limit of $(5.6)^{\varepsilon}$ is (4.4). So we expect $J^{\varepsilon,*}$ to converge to v ; the unique (viscosity) solution of (4.4). However the coefficients of $(5.6)^{\varepsilon}$ are not necessarily smooth, and the procedure of § 4 may not apply to this case. Still a convergence theorem holds if $(5.6)^{\varepsilon}$ has a comparison principle, which we define next.

DEFINITION 5.1. We say that $(5.6)^{\varepsilon}$ has a comparison principle if any viscosity subsolution w of $(5.6)^{\varepsilon}$ satisfying (4.1i) is less than or equal to any viscosity supersolution \tilde{w} of $(5.6)^{\varepsilon}$ satisfying (4.1i).

If, for example, $\mathcal{L}^{u^*,\xi}$ is as in (5.4), then $(5.6)^{\varepsilon}$ has a comparison principle for a large class of $u^*(\cdot)$. This class, in particular, includes the Lipschitz continuous functions.

We start our convergence proof with a lemma, which is due to Souganidis.

LEMMA 5.1. Suppose that the unique viscosity solutions v of (4.4) and v^{ε} of (3.1) are convex, continuously differentiable in the x -variable, and satisfy (4.1). Then $D_x v^{\varepsilon}(x, d, i)$ converges to $D_x v(x, d)$ uniformly on compact subsets of Σ , as $\varepsilon \rightarrow 0$.

Proof. Pick $\varepsilon_n \rightarrow 0$, $x_n \rightarrow x$ such that $p_n = D_x v^{\varepsilon_n}(x_n, d, i)$ converges to p . First, the convexity of $v^{\varepsilon_n}(\cdot, d, i)$ yields

$$v^{\varepsilon_n}(x_n + y, d, i) - v^{\varepsilon_n}(x_n, d, i) \geq p_n \cdot y, \quad \forall y.$$

Let n go to infinity to obtain

$$v(x + y, d) - v(x, d) \geq p \cdot y \quad \forall y.$$

Then the differentiability of v implies that $p = D_x v(x, d)$. \square

For $\xi = (x, d, i) \in \Sigma$, set

$$A^{\varepsilon}(\xi) = A(\xi; D_x v^{\varepsilon}(\xi), v^{\varepsilon}(x, \cdot, i); v^{\varepsilon}(x, d, \cdot)).$$

Equation (3.1) and the translation invariance (2.7) yield that

$$A^{\varepsilon}(\xi) = \frac{1}{\varepsilon} \left[v^{\varepsilon}(\xi) - \sum_{j \in Z} v^{\varepsilon}(x, d, j) \right].$$

Using the definition of A , we rewrite (3.1) as

$$(5.7) \quad v^{\varepsilon}(\xi) + H(\xi; D_x v^{\varepsilon}(\xi), v^{\varepsilon}(x, \cdot, i), A^{\varepsilon}(x, d, \cdot)) = 0.$$

Set

$$K^{\varepsilon}(\xi) = v^{\varepsilon}(\xi) - \mathcal{L}^{u^*(\xi),\xi}(D_x v^{\varepsilon}(\xi), v^{\varepsilon}(x, \cdot, i), A^{\varepsilon}(x, d, \cdot)) - G(\xi, u^*(\xi)).$$

In view of (5.3) and (5.7), $K^{\varepsilon}(\xi) \leq 0$.

LEMMA 5.2. Suppose that $(5.6)^{\varepsilon}$ has a comparison principle. Let J^{ε} be the viscosity solution of $(5.6)^{\varepsilon}$ satisfying (4.1). Then

$$(5.8) \quad v^{\varepsilon} \leq J^{\varepsilon}.$$

Proof. Since $K^{\varepsilon} \leq 0$, v^{ε} is a subsolution of $(5.6)^{\varepsilon}$. Therefore, the comparison principle yields $v^{\varepsilon} \leq J^{\varepsilon}$. \square

LEMMA 5.3. Assume the hypothesis of Theorem 4.1 and Lemma 5.1. Then K^ε converges to zero, as $\varepsilon \rightarrow 0$.

Proof. This follows from the continuity of A , (3.7), the convergence of v^ε and $D_x v^\varepsilon$, and the definition of \mathcal{L}^{u^*} . \square

The above result indicates that the difference $J^\varepsilon - v^\varepsilon$ should converge to zero. To prove this convergence, we assume that there are $\bar{K}, K, \nu, \bar{\nu} \geq 0$ such that for all $\xi \in \Sigma$,

$$(5.9) \quad |A(\xi; p, L; \alpha)| \leq \bar{K}(1 + |x|^{\bar{\nu}}) \quad \forall |p| + |L| + |\alpha| \leq K(1 + |x|^\nu),$$

and

$$(5.10) \quad \sup_{u \in K(\xi)} |\mathcal{L}^{u, \xi}(p, L, \kappa)| \leq \bar{K}(1 + |p| + |L| + |\kappa|) \quad \forall p \in R^n.$$

LEMMA 5.4. Assume (5.9), (5.10). Then for every $\tilde{K}, \tilde{\nu} \geq 0$ there is a continuously differentiable function $\eta(x) \geq 1$, such that

$$(5.11) \quad \frac{1}{2}\eta(x) - \mathcal{L}^{u, \xi}(D_x \eta(x), 0, 0) \geq 0 \quad \forall \xi \in \Sigma, \quad u \in K(\xi),$$

and

$$(5.12) \quad \eta(x) \geq \tilde{K}(1 + |x|^{\tilde{\nu}}).$$

Proof. Let $\eta(x) = C + \tilde{K}|x|^a$ for some $C \geq \tilde{K}$ and $a = \max\{2, \tilde{\nu}\}$. We will show that, for an appropriate choice of C , η satisfies (5.11) and (5.12). Indeed,

$$D_x \eta(x) = a\tilde{K}x|x|^{a-2},$$

and (5.10) yields

$$|\mathcal{L}^{u, \xi}(D_x \eta(x), 0, 0)| \leq \bar{K}(1 + [a\tilde{K}]|x|^{a-1}).$$

Therefore,

$$\frac{1}{2}\eta(x) - \mathcal{L}^{u, \xi}(D_x \eta(x), 0, 0) \geq (\frac{1}{2}C - \bar{K}) + \tilde{K}|x|^{a-1}(\frac{1}{2}|x| - a\bar{K}).$$

It is now elementary to show that the right-hand side of the above inequality is positive for every x if C is sufficiently large. \square

THEOREM 5.1. Assume the hypothesis of Theorem 4.1, Lemma 5.1, and (5.9), (5.10). Then $J^\varepsilon - v^\varepsilon$ converges to zero. In particular, u^* is asymptotically optimal.

Proof. Since v^ε satisfies (4.1), (5.9) implies that

$$|A^\varepsilon(\xi)| \leq \bar{K}(1 + |x|^{\bar{\nu}}).$$

In view of (5.10), the above estimate together with (4.1) yields

$$|K^\varepsilon(\xi)| \leq \tilde{K}(1 + |x|^{\tilde{\nu}-1})$$

for some $\tilde{K}, \tilde{\nu} \geq 1$. Let η be as in Lemma 5.4. Then (5.12) implies that $K^\varepsilon(\xi)/\eta(x)$ converges to zero uniformly on Σ , as $\varepsilon \rightarrow 0$. Set

$$k^\varepsilon = \inf \{|K^\varepsilon(\xi)/\eta(x)| : \xi \in \Sigma\}$$

and

$$w^\varepsilon(\xi) = v^\varepsilon(\xi) + 2k^\varepsilon \eta(x), \quad \xi = (x, d, i) \in \Sigma.$$

Then, the linearity of $\mathcal{L}^{u, \xi}$, the definition of K^ε , and (5.11) yield

$$\begin{aligned} w^\varepsilon(\xi) - \mathcal{L}^{u^*(\xi), \xi} \left(D_x w^\varepsilon(\xi), w^\varepsilon(x, \cdot, i), \frac{1}{\varepsilon} w^\varepsilon(x, d, \cdot) \right) - G(\xi, u^*(\xi)) \\ = K^\varepsilon(\xi) + 2k^\varepsilon [\eta(x) - \mathcal{L}^{u^*(\xi), \xi}(D_x \eta(x), 0, 0)] \\ \geq K^\varepsilon(\xi) + k^\varepsilon \eta(x) \geq 0. \end{aligned}$$

Hence w^ε is a supersolution of (5.6) ^{ε} . Consequently, the comparison principle implies that $J^\varepsilon \leq w^\varepsilon$. Now let ε go to zero and use the convergence of K^ε to zero together with (5.8) to obtain the convergence of J^ε to v . \square

Remark. If the operator $\mathcal{L}^{u^*(\xi),\xi}(p, L, k)$ is continuous in ξ , then $K^\varepsilon(\xi)$ converges to zero uniformly on compact subsets of Σ . Therefore w^ε , defined as in the above proof, converges to v uniformly on compact sets. Consequently, the conclusion of Theorem 5.1 holds uniformly on compact subsets of Σ .

Example 5.3. Consider the problem described in §2 with $n=2$, $m=1$, $D=\{(\delta, 1)\}$, $\gamma_1=\beta/K$, $\gamma_2=1/K$, $\mu_1(u)=\mu[\nu u_1+u_2]$, $\lambda_0=1$, and $G(x, y, u)=\alpha|x|+|y|$, where all the parameters are positive. Set

$$E = K\mu + 1,$$

$$\lambda = K\mu\nu + \beta.$$

Case 1. $\lambda \leq E\alpha$. In this case, the optimal feedback control is

$$u^*(x, y, 1) = \begin{cases} (0, 0) & \text{if } x > 0, y > 0, \\ (0, 1) & \text{if } x > 0, y = 0, \\ (0, K) & \text{if } x > 0, y < 0, \\ (E\delta, K - \lambda\delta)/[1 + \mu\delta(\beta - \nu)] & \text{if } x = 0, y < 0, \\ (K/\beta, 0) & \text{if } x < 0, \\ (\delta, 0) & \text{if } x = 0, y > 0, \\ (\delta, 1) & \text{if } x = 0, y = 0. \end{cases}$$

Case 2. $\lambda \geq E\alpha$. Then

$$u^*(x, y, 1) = \begin{cases} (0, 0) & \text{if } x > 0, y > 0, \\ (0, 1) & \text{if } x > 0, y = 0, \\ (0, K) & \text{if } y < 0, \\ (K - E, \lambda)/[\mu\nu + (1 - \mu)\beta] & \text{if } x < 0, y = 0, \\ (K/\beta, 0) & \text{if } x < 0, y > 0, \\ (\delta, 0) & \text{if } x = 0, y > 0, \\ (\delta, 1) & \text{if } x = 0, y = 0. \end{cases}$$

The value function is continuously differentiable in either case, and a comparison principle for (5.6) ^{ε} holds.

The above strategies differ only on the fourth quadrant. This is expected because in the other quadrants at most one of the products is in shortage, and then the optimal policy is to produce the product in shortage, if there is one, in full capacity. However, in the fourth quadrant of each of the products is in shortage, and therefore a priority rule is needed. The above calculations provide just this. In the first case, the optimal strategy is to produce x in full capacity. Hence in this case, the first product has priority over the second one. In the second case, this priority changes. So we may sum the above findings into the following rule:

If $\lambda \leq E\alpha$, produce the first product in full capacity if there is any shortage of it regardless the inventory level of the second product. If $\lambda \geq E\alpha$, reverse this rule.

6. Convergence rate. The exact rate at which $|v^\varepsilon - v|$ converges to zero is an interesting question. Recently Zhang and Sethi [15] obtained the rate $\varepsilon^{1/2}$ or $\varepsilon^{1/4}$ under different assumptions for the case described in Example 3.3. Also, they show, with an

explicit example, that in general $\varepsilon^{1/2}$ is the best rate. However, when the limit function is continuously differentiable we expect that the function $(v^\varepsilon(x, d, i) - v(x, d))/\varepsilon$ is uniformly bounded in ε on every compact subset of Σ . A similar result was proved in [7].

REFERENCES

- [1] R. AKELLA AND P. R. KUMAR, *Optimal control of production rate in a failure-prone manufacturing system*, IEEE Trans. Automat. Control, 13 (1986), pp. 116–126.
- [2] A. BENSOUSSAN, *Perturbation Methods in Optimal Control*, John Wiley, New York, 1988.
- [3] G. BITRAN AND D. TIRUPATI, *Hierarchical production planning*, 1989, preprint.
- [4] M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–501.
- [5] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc. 277 (1983), pp. 1–42.
- [6] W. H. FLEMING, S. P. SETHI, AND H. M. SONER, *An optimal stochastic production planning problem with randomly fluctuating demand*, SIAM J. Control Optim., 25 (1987), pp. 1494–1502.
- [7] W. H. FLEMING AND H. M. SONER, *Asymptotic expansions for Markov processes with Levy generators*, Appl. Math. Optim., 19 (1989), pp. 203–223.
- [8] S. GERSHWIN, *A Hierarchical Framework for Discrete Event Scheduling in Manufacturing Systems*, Lecture Notes in Control Inform. Sci., P. Varaiya and A. B. Kurzhanski, eds., Springer-Verlag, New York, 1987.
- [9] S. GERSHWIN, R. R. HILDEBRANT, R. SURI, AND S. K. MITTER, *Control perspective on recent trends in manufacturing systems*, IEEE Control Systems Mag., 6 (1986), pp. 3–16.
- [10] H. J. KUSHNER, *Diffusion approximations and nearly optimal policies for system breakdown and repair problem*, Appl. Math. Optim., 20 (1989), pp. 33–53.
- [11] J. P. LEHOCZKY, S. P. SETHI, H. M. SONER, AND M. TAKSAR, *An asymptotic analysis of hierarchical control of manufacturing systems*, Math. Oper. Res., 16 (1991), pp. 596–608.
- [12] P.-L. LIONS, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, Boston, 1982.
- [13] V. R. SAKSENA, J. O'REILLY, AND P. V. KOKOTOVIC, *Singular perturbations and time-scale methods in control theory: Survey 1976–1983*, Automatica, 20 (1984), pp. 273–293.
- [14] H. M. SONER, *Optimal control with state space constraint II*, SIAM J. Control Optim., 24 (1986), pp. 1110–1122.
- [15] Q. ZHANG AND S. P. SETHI, *Hierarchical production planning in dynamic stochastic manufacturing systems: Asymptotic optimality and error bounds*, preprint.

MAXIMUM LIKELIHOOD ESTIMATOR FOR TWO-POINT BOUNDARY VALUE PROCESS*

SHIN ICHI AIHARA† AND ARUNABHA BAGCHI‡

Abstract. The problem of identifying unknown parameters in two-point boundary value (TPBV) processes is studied. Using the explicit form of the likelihood functional, the consistency property of the maximum likelihood estimator is analyzed under a large number of independent experiments.

Key words. two-point boundary value (TPBV) process, maximum likelihood estimator, consistency, likelihood functional

AMS(MOS) subject classifications. 93E12, 60H99, 62M09

1. Introduction. The smoothing problem for boundary value (TPBV) processes has been extensively studied by Adams, Willsky, and Levy [1], [2]. These are random fields (processes) satisfying partial (ordinary) differential equations with noisy boundary conditions. A rigorous mathematical treatment of boundary value processes and the related smoothing and identification problems may be found in Bagchi and Aihara [3]. The likelihood functional derived in that paper is not given in terms of system parameters and is therefore not particularly suitable for identification purposes. In Bagchi and Westdijk [4], this likelihood formula has been further worked out for the one-parameter case, the so-called two-point boundary value (TPBV) processes, and expressed in terms of the system parameters. In this paper, we use this expression to study consistency of the maximum likelihood estimates of the unknown system parameters for TPBV processes.

TPBV processes are observed in a fixed-time interval; say $[0, T]$. Based on one realization of the observation process in this time interval, no statement can be made about the asymptotic properties of estimates of the unknown parameters. In this paper, we make n independent experiments. Based on the observations $y_t^1, y_t^2, \dots, y_t^n, 0 \leq t \leq T$, we obtain the maximum likelihood estimate of the unknown system parameters. Using the approach proposed by Borkar and Bagchi [5], we show the consistency of the maximum likelihood estimate obtained for TPBV processes.

2. Problem formulation and the likelihood functional. Suppose that $\{x_t, t \geq 0\}$ is an n -dimensional, real-valued, time-invariant, TPBV process satisfying

$$(2.1) \quad dx_t = A(\theta)x_t dt + B(\theta) dw_t$$

and the boundary condition

$$(2.2) \quad V^0 x_0 + V^T x_T = v,$$

where $\{w_t, t \geq 0\}$ is a p -dimensional standard Brownian motion. v is an n -dimensional Gaussian random vector, independent of $\{w_t, 0 \leq t \leq T\}$, with $Ev = 0$ and $Evv^* = \Sigma_v > 0$, with $(*)$ denoting transpose.

* Received by the editors November 5, 1990; accepted for publication (in revised form) November 13, 1991.

† Department of Management and System Science, The Science University of Tokyo, Suwa College, 5000-1 Toyohira, Chino, Nagano 391-02, Japan.

‡ Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands.

$A(\theta)$ is an $n \times n$ matrix and $B(\theta)$ is an $n \times p$ matrix. The components of $A(\theta)$ and $B(\theta)$ depend on the unknown parameter vector $\theta \in \Theta$, where Θ is a compact set in \mathbb{R}^k . V^0 and V^T are constant $n \times n$ matrices.

The measurement process $\{y_t, 0 \leq t \leq T\}$ is given by

$$(2.3) \quad dy_t = Cx_t dt + D dw_t, \quad y_0 = 0,$$

where y_t is an m -dimensional random vector and the matrices C and D are known. We assume that $BD^* = 0$ (the state and measurement noises are independent) and that DD^* is a positive definite matrix. Without loss of generality, we may take $DD^* = I$ (identity matrix).

Our problem is to determine an estimate for θ based on n independent experiments that give us a sample $y_t^1, \dots, y_t^n, 0 \leq t \leq T$ of n independent trajectories of the observation process $\{y_t, 0 \leq t \leq T\}$. Based on these samples, we determine the maximum likelihood estimate $\hat{\theta}_n$ of the unknown parameter θ by maximizing the likelihood functional for the problem. Our objective is to show consistency of the maximum likelihood estimator obtained for the unknown parameter vector θ .

We derive in this section the exact form of the likelihood functional for our problem. The detailed derivation may be found in [4] and is based on an important result of Shepp [6]. We present here only the broad outline of the results obtained in [4], which helps us also in setting the notations used throughout the paper. It is easy to solve (2.1), (2.2), and the solution is given by

$$(2.4) \quad x_t = \Phi(t)(V^0 + V^T \Phi(T))^{-1} \cdot \left\{ V^0 \int_0^t \Phi(u)^{-1} B dw_u - V^T \Phi(T) \int_t^T \Phi(u)^{-1} B dw_u + v \right\},$$

where the (fixed) parameter θ is not explicitly mentioned, the matrix $V^0 + V^T \Phi(T)$ is assumed to be invertible, and $\Phi(t) \equiv e^{At}$, the state transition matrix for the system. The signal process in (2.3) is then

$$(2.5) \quad Cx_t = V(t)v + \int_0^T h(t, u) dw_u, \quad 0 \leq t \leq T,$$

where

$$(2.6a) \quad V(t) = C\Phi(t)(V^0 + V^T \Phi(T))^{-1}$$

and

$$(2.6b) \quad h(t, u) = \begin{cases} V(t)V^0\Phi(u)^{-1}B, & 0 \leq u \leq t, \\ -V(t)V^T\Phi(T)\Phi(u)^{-1}B, & t < u \leq T. \end{cases}$$

Define

$$(2.7) \quad \begin{aligned} r(t, u) &\triangleq E\{(Cx_t)(Cx_u)^*\} \\ &= V(t)\Sigma_v V(u)^* + \int_0^T h(t, z)h(u, z)^* dz. \end{aligned}$$

Define the operator $R: L_2^m[0, T] \rightarrow L_2^m[0, T]$ by

$$(2.8) \quad (Rf)(t) = \int_0^T r(t, u)f(u) du.$$

The covariance operator R_y of the observation process is then given by

$$(2.9) \quad R_y = I + R.$$

Noting that the operator R is trace class, the Fredholm determinant is

$$(2.10) \quad d = \prod_j (1 + \lambda_j),$$

where λ_j 's are the eigenvalues of R . Furthermore, there exists a unique operator K such that

$$(2.11) \quad K = R(I + R)^{-1} = I - (I + R)^{-1}.$$

Let p_y denote the measure induced by the process y_t , $0 \leq t \leq T$ on the space $C([0, T]; \mathbb{R}^m)$ of \mathbb{R}^m -valued continuous functions on $[0, T]$ with p_w denoting the Wiener measure thereon. By a result of Shepp [6] mentioned above, p_y is absolutely continuous with respect to p_w , and the Radon-Nikodym derivative is given by

$$(2.12) \quad \frac{dp_y}{dp_w}(y) = d^{-1/2} \exp \left\{ \frac{1}{2} \int_0^T \left[\int_0^T k(t, s) dy_s, dy_t \right] \right\},$$

where $k(t, s)$ is the kernel of the operator K and the integral on the right-hand side of (2.12) is a *double Wiener integral* as defined in [6].

It may be possible to define this integral as a quasimartingale following Itô [7], but in that case a correction term would appear in the right-hand side of (2.12).

Using Krein factorization [8], we get

$$(2.13) \quad R_y^{-1} = (I + R)^{-1} = (I - L^*)(I - L),$$

where L is a Volterra operator with kernel $l(t, s)$ and, as is shown in [8, pp. 232-234],

$$(2.14) \quad \begin{aligned} \log d^{-1/2} &= \frac{1}{2} \text{Tr} \log (I + R)^{-1} = -\frac{1}{2} \text{Tr} (L + L^*) \\ &= -\frac{1}{2} \int_0^T \text{Tr} l(t, t) dt. \end{aligned}$$

It has been shown in [4] that

$$(2.15) \quad \int_0^T \text{Tr} l(t, t) dt = \int_0^T \text{Tr} [C \{F(t)^{-1}\}_{1,2} C^*] dt,$$

where $\{F(t)^{-1}\}_{1,2}$ is the $(1, 2)$ th block of the block matrix $F(t)^{-1}$ with

$$(2.16) \quad F(t) \triangleq W^0(t) e^{-\Delta t} + W^T(t).$$

Here is

$$(2.17) \quad W^0(t) \triangleq \begin{pmatrix} \Pi_0(t) & -I \\ \Pi_1(t)^* & 0 \end{pmatrix}, \quad W^T(t) \triangleq \begin{pmatrix} \Pi_1(t) & 0 \\ \Pi_2(t) & I \end{pmatrix}, \quad \Delta \triangleq \begin{pmatrix} A & BB^* \\ C^*C & -A^* \end{pmatrix},$$

and $\Pi_i(t)$, $i = 0, 1, 2$, satisfy the differential equations

$$(2.18a) \quad \dot{\Pi}_0(t) = \Pi_1(t) BB^* \Pi_1(t)^*, \quad \Pi_0(T) = V^0 \Sigma_v^{-1} V^0,$$

$$(2.18b) \quad \dot{\Pi}_1(t) = \Pi_1(t) BB^* \Pi_2(t) - \Pi_1(t) A, \quad \Pi_1(T) = V^{0*} \Sigma_v^{-1} V^T,$$

$$(2.18c) \quad \dot{\Pi}_2(t) = \Pi_2(t) BB^* \Pi_2(t) - \Pi_2(t) A - A^* \Pi_2(t), \quad \Pi_2(T) = V^{T*} \Sigma_v^{-1} V^T.$$

Furthermore, as shown in [4], with \mathcal{Y}_t denoting the smallest σ -algebra generated by y_s , $0 \leq s \leq t$,

$$(2.19) \quad C \hat{x}_t \triangleq CE[x_2 | \mathcal{Y}_T] = \int_0^T k(t, s) dy_s.$$

Therefore, we get

$$(2.20) \quad \log \frac{dp_y}{dp_w}(y) = \frac{1}{2} \int_0^T [C\hat{x}_t, dy_t] - \frac{1}{2} \int_0^T \text{Tr} [C\{F(t)^{-1}\}_{1,2}C^*] dt.$$

Suppose now that we make n independent experiments, yielding independent sample trajectories $y_t^1, y_t^2, \dots, y_t^n$. For $k = 1, 2, \dots, n$, let

$$(2.21) \quad dx_t^k(\theta) = A(\theta)x_t^k(\theta) dt + B(\theta) dw_t^k, \quad V^0 x_0^k + V^T x_T^k = v^k,$$

and

$$(2.22) \quad dy_t^k = Cx_t^k dt + D dw_t^k.$$

Let $\hat{x}_t^k \triangleq E[x_t^k | \mathcal{Y}_t^k]$.

Then the log likelihood functional is given by

$$(2.23) \quad \begin{aligned} L_n(y^1, \dots, y^n; \theta) &= \log \prod_{k=1}^n \frac{dp_y}{dp_w}(y^k) = \sum_{k=1}^n \log \frac{dp_y}{dp_w}(y^k) \\ &= \frac{1}{2} \sum_{k=1}^n \left(\int_0^T [C\hat{x}_t^k(\theta), dy_t^k] - \int_0^T \text{Tr} [C\{F(t; \theta)^{-1}\}_{1,2}C^*] dt \right). \end{aligned}$$

3. Sample properties of signal and observation processes. We begin with the strong laws for independent sample trajectories.

THEOREM 3.1 (strong law of large numbers). *It holds that*

$$(3.1) \quad \frac{1}{n} \sum_{k=1}^n Cx_t^k \rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \forall t \in [0, T] \text{ a.s.}$$

and

$$(3.2) \quad \frac{1}{n} \sum_{k=1}^n (Cx_t^k)(Cx_t^k)^* \rightarrow E\{Cx_t(Cx_t)^*\} \quad \text{as } n \rightarrow \infty \quad \forall t \in [0, T] \text{ a.s.}$$

with $\{x_t, 0 \leq t \leq T\}$ given by (2.4).

Proof. From the strong law of large numbers in [9, p. 363], to prove (3.1), we only need to check that

$$E\{|Cx_t^k|^4\} \leq \text{Const independent of } k.$$

Noting that Cx_t^k is a Gaussian random process, we have

$$E\{|Cx_t^k|^4\} = 3(E\{|Cx_t^k|^2\})^2.$$

It follows from (2.5) that

$$\begin{aligned} E\{|Cx_t^k|^4\} &= 3 \left(\text{Tr} \{V(t)\Sigma_v V(t)^*\} + \int_0^T \text{Tr} \{h(t, u)h(t, u)^*\} du \right)^2 \\ &\leq \text{Const independent of } k. \end{aligned}$$

The proof of (3.2) is similar. \square

THEOREM 3.2. *It holds that*

$$(3.3) \quad \begin{aligned} &\frac{1}{n} \sum_{k=1}^n \int_0^T \left[\int_0^T k(t, s) dy_s^k, dy_t^k \right] \\ &= \frac{1}{n} \sum_{k=1}^n \int_0^T [C\hat{x}_t^k, dy_t^k] \\ &\rightarrow E \int_0^T [C\hat{x}_t, Cx_t] dt + \text{Tr } K \text{ a.s. as } n \rightarrow \infty. \end{aligned}$$

Proof. Let

$$(3.4) \quad I_n \triangleq \frac{1}{n} \sum_{k=1}^n \int_0^T \left[\int_0^T k(t, s) dy_s^k, dy_t^k \right].$$

Then

$$(3.5) \quad \begin{aligned} I_n &= \frac{1}{n} \left(\sum_{k=1}^n \int_0^T [C\hat{x}_t^k, Cx_t^k] dt + \sum_{k=1}^n \int_0^T \left[\int_0^T k(t, s) Cx_s^k ds, D dw_t^k \right] \right. \\ &\quad \left. + \sum_{k=1}^n \int_0^T \left[\int_0^T k(t, s) D dw_s^k, D dw_t^k \right] \right) \\ &= I_n^1 + I_n^2 + I_n^3, \end{aligned}$$

say. Using the same method used in Theorem 3.1, we have

$$(3.6) \quad I_n^2 \rightarrow E \int_0^T \left[\int_0^T k(t, s) Cx_s ds, D dw_t \right] = 0 \quad \text{a.s. as } n \rightarrow \infty,$$

and the independence of the signal and measurement noises implies that

$$(3.7) \quad \begin{aligned} I_n^1 &= \frac{1}{n} \sum_{k=1}^n \left(\int_0^T \left[\int_0^T k(t, s) Cx_s^k ds, Cx_t^k \right] dt \right. \\ &\quad \left. + \int_0^T \left[\int_0^T k(t, s) D dw_s^k, Cx_t^k \right] dt \right) \\ &\rightarrow E \int_0^T \int_0^T [k(t, s) Cx_s, Cx_t] ds dt \\ &= E \int_0^T [C\hat{x}_t, Cx_t] dt \quad \text{a.s. as } n \rightarrow \infty. \end{aligned}$$

All we must do now is to study the convergence of I_n^3 , which is defined in the sense of a double Wiener integral.

From Shepp [6], we have

$$E(I_n^3) = \frac{1}{n} \sum_{k=1}^n \int_0^T \text{Tr } k(t, t) dt = \text{Tr } K.$$

Let $J_k = \int_0^T [\int_0^T k(t, s) D dw_s^k, D dw_t^k] - \text{Tr } K$. Then $\{J_k\}$ is a sequence of mutually independent random vectors of zero mean, and, by the strong law of large numbers,

$$(3.8) \quad I_n^3 \rightarrow \text{Tr } K \quad \text{a.s. as } n \rightarrow \infty.$$

Combining (3.6)–(3.8), we get the desired result. \square

4. The identification problem. We have already obtained the likelihood functional $L(y^1, \dots, y^n; \theta)$ for our problem. Since Θ is a compact set, for each ω in the sample space, we can find an element $\hat{\theta}_n(\omega) \in \Theta$ such that

$$(4.1) \quad L(y^1, \dots, y^n; \hat{\theta}_n) \geq L(y^1, \dots, y^n; \theta') \quad \text{for all } \theta' \in \Theta;$$

that is,

$$\sum_{k=1}^n \log \frac{dp_y^\theta}{dp_w}(y^k) \big|_{\theta=\hat{\theta}_n} \geq \sum_{k=1}^n \log \frac{dp_y^{\theta'}}{dp_w}(y^k) \quad \text{for all } \theta' \in \Theta.$$

In particular, for $\theta' = \theta_0$ (the true value of the parameter, which is assumed to lie in Θ), we have

$$(4.2) \quad \sum_{k=1}^n \log \frac{dp_y^\theta}{dp_{y_0}^{\theta_0}}(y^k) \geq 0.$$

We call $\hat{\theta}_n$ the *maximum likelihood estimate* (MLE) of θ_0 based on n independent sample trajectories y^1, \dots, y^n .

To establish consistency of the MLE, we must check the continuity of $\hat{x}_t^k(\theta)$ with respect to θ . For this, we need the following assumption.

Assumption A1. There exists a constant C such that

$$\begin{aligned} \sup_{\theta \in \Theta} \|A(\theta)\| &\leq C, & \sup_{\theta \in \Theta} \|\nabla_\theta A(\theta)\| &\leq C, \\ \sup_{\theta \in \Theta} \|B(\theta)\| &\leq C, & \sup_{\theta \in \Theta} \|\nabla_\theta B(\theta)\| &\leq C. \end{aligned}$$

LEMMA 4.1. *It holds that*

$$(4.3) \quad \begin{aligned} \lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \frac{1}{n} \sum_{k=1}^n \int_0^T [C\hat{x}_t^k(\theta), D dw_t^k] \\ = \sup_{\theta \in \Theta} \int_0^T \text{Tr } k(t, t; \theta) dt \quad a.s. \end{aligned}$$

Proof. The main idea of this proof is to transform the double Wiener integral into the Itô type integral. Then we can directly apply the results proposed by Borkar and Bagchi [5] to our case. It is easy to show that

$$(4.4) \quad \begin{aligned} \frac{1}{n} \sum_{k=1}^n \int_0^T [C\hat{x}_t^k(\theta), D dw_t^k] &= \frac{1}{n} \sum_{k=1}^n \int_0^T \left[\int_0^T k(t, s; \theta) dy_s^k, D dw_t^k \right] \\ &= \frac{1}{n} \sum_{k=1}^n \int_0^T \left(\int_0^T k(t, s; \theta) Cx_t^k(\theta_0) ds \right)^* D dw_t^k \\ &\quad + \frac{1}{n} \sum_{k=1}^n \int_0^T \left[\int_0^T k(t, s; \theta) D dw_s^k, D dw_t^k \right], \end{aligned}$$

where the first term of the right-hand side of (4.4) is defined in the Itô sense [7] because Cx^k and Dw^k are independent, and we find that Dw_t^k is \mathcal{F}_t -martingale for $\mathcal{F}_t = \sigma\{Dw_s^k; 0 \leq s \leq T\}$. Using the integration by parts formula for the double Wiener integral, the second term of the right-hand side of (4.4) becomes

$$(4.5) \quad \begin{aligned} &\frac{1}{n} \sum_{k=1}^n \int_0^T \left[\int_0^T k(t, s; \theta) D dw_s^k, D dw_t^k \right] \\ &= -\frac{1}{n} \sum_{k=1}^n w_T^{k*} D^* k(T, T; \theta) D w_T^k \\ &\quad + \frac{1}{n} \sum_{k=1}^n w_T^{k*} D^* \int_0^T \{k(t, T; \theta) + k(T, t; \theta)\} D dw_t^k \\ &\quad + \frac{1}{n} \sum_{k=1}^n \int_0^T \int_0^T w_t^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D w_s^k ds dt \\ &= I_1^n + I_2^n + I_3^n, \end{aligned}$$

say. With the aid of Itô's formula and $DD^* = I$, we have

$$\begin{aligned} I_1^n &= -\frac{2}{n} \sum_{k=1}^n \int_0^T w_t^{k*} D^* k(T, T; \theta) D dw_t^k - \text{Tr} \{k(T, T; \theta)\} \cdot T, \\ I_2^n &= \frac{2}{n} \sum_{k=1}^n \int_0^T w_t^{k*} D^* (k(t, T; \theta) + k(T, t; \theta)) D dw_t^k \\ &\quad + \int_0^T \text{Tr} \{k(t, T; \theta) + k(T, t; \theta)\} dt, \end{aligned}$$

and, for the integrand of I_3^n term, setting $t > s$,

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^n w_t^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D w_s^k &= \frac{1}{n} \sum_{k=1}^n w_s^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D w_s^k \\ &\quad + \frac{1}{n} \sum_{k=1}^n \int_s^T w_s^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D dw_t^k; \end{aligned}$$

applying the Itô formula to the first term, we have

$$\begin{aligned} &= \frac{1}{n} \sum_{k=1}^n \left\{ \int_0^s 2w_\tau^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D dw_\tau^k + \int_s^t w_s^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D dw_\tau^k \right\} \\ &\quad + \text{Tr} \left\{ \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} \right\}_s. \end{aligned}$$

Hence

$$\begin{aligned} I_3^n &= \frac{1}{n} \sum_{k=1}^n \int_0^T \int_0^T \left\{ \int_0^{s \wedge t} 2w_\tau^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D dw_\tau^k \right. \\ &\quad \left. + \int_{s \wedge t}^{s \vee t} w_s^{k*} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D dw_\tau^k \right\} dt ds \\ &\quad + \int_0^T \int_0^T \text{Tr} \left\{ \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} \right\} (t \wedge s) dt ds. \end{aligned}$$

Combining all estimates, we have

$$\begin{aligned} &\frac{1}{n} \sum_{k=1}^n \int_0^T [C\hat{x}_t^k(\theta), D dw_t^k] \\ &= \frac{1}{n} \sum_{k=1}^n \int_0^T f^k(t, w^k; \theta) dw_t^k \\ (4.6) \quad &+ \int_0^T \int_0^T \frac{1}{n} \sum_{k=1}^n \int_0^{s \vee t} g^k(t, s, \tau, w^k) dw_\tau^k ds dt \\ &- \text{Tr} \{k(T, T; \theta)\} + \int_0^T \text{Tr} \{k(t, T; \theta) + k(T, t; \theta)\} dt \\ &+ \int_0^T \int_0^T \text{Tr} \left\{ \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} \right\} (t \wedge s) dt ds, \end{aligned}$$

where

$$\begin{aligned} f^k(t, w^k; \theta) &= \int_0^T k(t, s; \theta) Cx_s^k(\theta_0) ds D - 2w_t^{k*} D^* k(T, T; \theta) D \\ &\quad + 2w_t^{k*} D^* (k(t, T; \theta) + k(T, t; \theta)) D \end{aligned}$$

and

$$g^k(t, s, \tau, w^k) = \{2\chi_{s < t} w_\tau^{k*} + (1 - \chi_{s < t}) w_s^{k*}\} D^* \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} D,$$

where

$$\chi_{s < t} = \begin{cases} 1 & \text{if } s < t, \\ 0 & \text{if } s > t. \end{cases}$$

Hence, we must prove that

$$(4.7) \quad \lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{k=1}^n \int_0^T f^k(t, w^k; \theta) dw_t^k \right) = 0 \quad \text{a.s.}$$

and

$$(4.8) \quad \lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \left(\frac{1}{n} \sum_{k=1}^n \int_0^{s \vee t} g^k(t, s, \tau; w^k) dw_\tau^k \right) = 0 \quad \text{a.s.}$$

To apply the results of Borkar and Bagchi [5], we define

$$\tilde{w}_t = [w_t^1, w_t^2, \dots, w_t^n]^*$$

and

$$\tilde{f}(t, \tilde{w}, \theta) = [f^1(t, w^1; \theta), f^2(t, w^2; \theta), \dots, f^n(t, w^n; \theta)]$$

so that

$$(4.9) \quad \frac{1}{n} \sum_{k=1}^n \int_0^T f^k(t, w^k; \theta) dw_t^k = \frac{1}{n} \int_0^T \tilde{f}(t, \tilde{w}, \theta) d\tilde{w}_t.$$

Hence, for every $\theta_1, \theta_2 \in \Theta$, $2m \geq p+1$ ($\Theta \subset R^p$), and $m \geq 2$, by using the same argument as Lemma 3.1 of [5, p. 196], we have

$$\begin{aligned} & E \left\{ \left(\frac{1}{n} \int_0^T (\tilde{f}(t, \tilde{w}; \theta_1) - \tilde{f}(t, \tilde{w}; \theta_2)) d\tilde{w}_t \right)^{2m} \right\} \\ & \leq (m(2m-1))^{m-1} T^{m-1} \frac{1}{n^{2m}} \int_0^T E \left\{ \left(\sum_{k=1}^n \|f^k(t, w^k; \theta_1) - f^k(t, w^k; \theta_2)\|^2 \right)^m \right\} dt \\ & \leq C(m) \frac{n \cdot m}{n^{2m}} \sum_{k=1}^n \int_0^T E \{ \|f^k(t, w^k; \theta_1) - f^k(t, w^k; \theta_2)\|^{2m} \} dt, \end{aligned}$$

where

$$\begin{aligned} C(m) &= (m(2m-1))^{m-1} T^{m-1} \\ & \leq C(m) m \frac{n}{n^{2m}} \sum_{k=1}^n \int_0^T E \left\{ \left\| \int_0^T (k(t, s; \theta_1) - k(t, s; \theta_2)) Cx_s^k(\theta_0) ds \right. \right. \\ & \quad - 2w_t^{k*} D^* (k(T, T; \theta_1) - k(T, T; \theta_2)) D \\ & \quad + 2w_t^{k*} D^* \{ (k(t, T; \theta_1) - k(t, T; \theta_2)) \\ & \quad \left. \left. + (k(T, t; \theta_1) - k(T, t; \theta_2)) \} D \right\|^{2m} \right\} dt. \end{aligned} \quad (4.10)$$

From Assumption A1, we can show that

$$\|k(t, s; \theta_1) - k(t, s; \theta_2)\| \leq C \|\theta_1 - \theta_2\| \quad \text{for } 0 \leq s, \quad t \leq T.$$

Furthermore,

$$E \left\{ \left(\int_0^T \|x_s^k(\theta_0)\| ds \right)^{2m} \right\} \leq \text{Const independent of } k$$

and

$$E \{ \|w_t^k\|^{2m} \} \leq \text{Const independent of } k.$$

Hence, (4.10) becomes

$$\begin{aligned} E \left\{ \left(\frac{1}{n} \int_0^T (\tilde{f}(t, \tilde{w}; \theta_1) - \tilde{f}(t, \tilde{w}; \theta_2)) d\tilde{w}_t \right)^{2m} \right\} \\ \leq \text{Const} \frac{n^2}{n^{2m}} \|\theta_1 - \theta_2\|^{2m}, \end{aligned}$$

noting that, from $m \geq 2$, there exists an $\alpha > 0$ such that

$$(4.11) \quad \leq \text{Const} \frac{\|\theta_1 - \theta_2\|^{2m}}{n^{1+\alpha}}.$$

This estimate is the same as Lemma 3.1 [5, p. 196]; hence (4.7) can be derived. Repeating the same argument as above, we can prove (4.8). Details are omitted.

From (4.6)–(4.8), we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \sup_{\theta \in \Theta} \frac{1}{n} \int_0^T [C\hat{x}_t^k(\theta), D dw_s^k] \\ = \sup_{\theta \in \Theta} \left\{ -\text{Tr} \{k(T, T; \theta)\} + \int_0^T \text{Tr} \{k(t, T; \theta) + k(T, t; \theta)\} dt \right. \\ \left. + \int_0^T \int_0^T \text{Tr} \left\{ \frac{\partial^2 k(t, s; \theta)}{\partial t \partial s} \right\} (t \wedge s) dt ds \right\} \quad \text{a.s.} \end{aligned}$$

(integrating by parts)

$$= \sup_{\theta \in \Theta} \int_0^T \text{Tr} k(t, t; \theta) dt \quad \text{a.s.}$$

This is the desired result. \square

THEOREM 4.1. *Let \mathcal{M} be the set of measure zero outside of which Lemma 4.1 holds. For each $\omega \notin \mathcal{M}$, letting $\hat{\theta}_n$ be the maximum likelihood estimate of θ_0 , we have the following strong consistency property:*

$$(4.12) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \int_0^T |C(\hat{x}_t^k(\hat{\theta}_n) - x_t^k(\theta_0))|^2 dt = \int_0^T \text{Tr} [CP(t; \theta_0)C^*] dt \quad \text{a.s.}$$

Proof. From (4.2), the MLE $\hat{\theta}_n$ satisfies

$$(4.13) \quad \left(\frac{d(1, \theta_0)}{d(1, \hat{\theta}_n)} \right)^{n/2} \exp \left\{ \frac{1}{2} \sum_{k=1}^n \int_0^T [C(\hat{x}_t^k(\hat{\theta}_n) - \hat{x}_t^k(\theta_0)), dy_t^k] \right\} \geq 1.$$

From (2.14) and (2.15), we find that

$$\begin{aligned} \log \left(\frac{d(1, \theta_0)}{d(1, \hat{\theta}_n)} \right)^{n/2} &= \frac{n}{2} \{ \text{Tr} \{K(\theta_0) - K(\hat{\theta}_n)\} \} \\ &= \frac{n}{2} \int_0^T \text{Tr} \{CP(t; \theta_0)C^* - CP(t; \hat{\theta}_n)C^*\} dt. \end{aligned}$$

Hence, from (4.13),

$$(4.14) \quad \begin{aligned} & \frac{1}{2} \sum_{k=1}^n \int_0^T [C(\hat{x}_t^k(\hat{\theta}_n) - \hat{x}_t^k(\theta_0)), dy_t^k] \\ & - \frac{n}{2} \left\{ \int_0^T \text{Tr}[CP(t; \hat{\theta}_n)C^* - CP(t; \theta_0)C^*] dt \right\} \geq 0. \end{aligned}$$

Dividing (4.14) by n , the first term of the above inequality becomes

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T [C(\hat{x}_t^k(\hat{\theta}_n) - \hat{x}_t^k(\theta_0)), dy_t^k] \\ & = \lim_{n \rightarrow \infty} \left\{ \frac{1}{2n} \sum_{k=1}^n \int_0^T [C(\hat{x}_t^k(\hat{\theta}_n) - \hat{x}_t^k(\theta_0)), Cx_t^k(\theta_0)] dt \right. \\ & \quad \left. + \frac{1}{2n} \sum_{k=1}^n \int_0^T [C(\hat{x}_t^k(\hat{\theta}_n) - \hat{x}_t^k(\theta_0)), D dw_t^k] \right\} \\ & = \lim_{n \rightarrow \infty} \{T_1^n + T_2^n\}, \end{aligned}$$

say. From Lemma 4.1 and (3.3), it follows that

$$(4.15) \quad \begin{aligned} \lim_{n \rightarrow \infty} T_2^n &= \lim_{n \rightarrow \infty} \frac{1}{2} \int_0^T \text{Tr}[k(t, t; \hat{\theta}_n) - k(t, t; \theta_0)] dt \\ &= \lim_{n \rightarrow \infty} \frac{1}{2} \int_0^T \text{Tr}[CP(t; \hat{\theta}_n)C^* - CP(t; \theta_0)C^*] dt \quad \text{a.s.} \end{aligned}$$

Hence, dividing (4.14) by n and using (4.15), we have

$$(4.16) \quad \lim_{n \rightarrow \infty} T_1^n \geq 0 \quad \text{a.s.}$$

This implies that

$$(4.17) \quad \begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T [Cx_t^k(\hat{\theta}_n), Cx_t^k(\theta_0)] dt \\ & \geq \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T [C\hat{x}_t^k(\theta_0), Cx_t^k(\theta_0)] dt \\ & = E \left\{ \int_0^T [C\hat{x}_t(\theta_0), Cx_t(\theta_0)] dt \right\} \\ & = E \left\{ \int_0^T |C\hat{x}_t(\theta_0)|^2 dt \right\} \quad \text{a.s.} \end{aligned}$$

Hence, by using (4.17), the following inequality can be derived:

$$(4.18) \quad \begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T [C(x_t^k(\theta_0) - C\hat{x}_t^k(\hat{\theta}_n)), Cx_t^k(\theta_0)] dt \\ & \leq \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T |C(x_t^k(\theta_0)|^2 dt - E \left\{ \int_0^T |C\hat{x}_t(\theta_0)|^2 dt \right\} \\ & = E \left\{ \int_0^T |Cx_t(\theta_0) - C\hat{x}_t(\theta_0)|^2 dt \right\} \\ & = \int_0^T \text{Tr}[CP(t; \theta_0)C^*] dt \quad \text{a.s.} \end{aligned}$$

On the other hand, from (4.18), we also have

$$\begin{aligned}
 & \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T [C(x_t^k(\theta_0) - C\hat{x}_t^k(\hat{\theta}_n)), Cx_t^k(\theta_0)] dt \\
 &= \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \left\{ \int_0^T |Cx_t^k(\theta_0)|^2 dt - \int_0^T |C\hat{x}_t^k(\hat{\theta}_n)|^2 dt \right\} \\
 &+ \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T [C(\hat{x}_t^k(\hat{\theta}_n), C\hat{x}_t^k(\hat{\theta}_n) - Cx_t^k(\theta_0))] dt \\
 &\leq \int_0^T \text{Tr}[CP(t; \theta_0)C^*] dt \quad \text{a.s.}
 \end{aligned}$$

Then

$$\begin{aligned}
 & \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \int_0^T [C(\hat{x}_t^k(\hat{\theta}_n), C\hat{x}_t^k(\hat{\theta}_n) - Cx_t^k(\theta_0))] dt \\
 &\leq \int_0^T \text{Tr}[CP(t; \theta_0)C^*] dt \\
 &- \lim_{n \rightarrow \infty} \frac{1}{2n} \sum_{k=1}^n \left\{ \int_0^T |Cx_t^k(\theta_0)|^2 dt - \int_0^T |C\hat{x}_t^k(\hat{\theta}_n)|^2 dt \right\} \\
 (4.19) \quad &= \int_0^T \text{Tr}[CP(t; \theta_0)C^*] dt \\
 &- \lim_{n \rightarrow \infty} E \left\{ \int_0^T |Cx_t(\theta_0)|^2 dt - \int_0^T |C\hat{x}_t(\hat{\theta}_n)|^2 dt \right\} \\
 &\leq 0 \quad \text{a.s.},
 \end{aligned}$$

where we used the fact that $P(t; \theta_0)$ is an error covariance in the sense of minimum variance estimate, i.e.,

$$\text{Tr}\{P(t, \theta_0)\} - E\{|x_t(\theta)|^2 - |\hat{x}_t(\theta)|^2\} \leq 0 \quad \forall \theta \in \theta.$$

Then, summing (4.18) and (4.19), (4.12) can be derived. \square

COROLLARY. *It holds that*

$$\hat{\theta}_n \rightarrow \{\theta \mid P_{\theta_0}\{C\hat{x}_t^n(\theta) = C\hat{x}_t^n(\theta_0) \text{ a.e.t., } \forall n\} = 1\} \quad \text{a.s.}$$

Proof. The proof follows from the previous theorem and the fact that

$$\begin{aligned}
 & \frac{1}{n} \sum_{k=1}^n \int_0^T |C\hat{x}_t^k(\theta) - Cx_t^k(\theta_0)|^2 dt \rightarrow E \int_0^T |C\hat{x}_t(\theta) - C\hat{x}_t(\theta_0)|^2 dt \\
 &+ \int_0^T \text{Tr}(CP(t; \theta_0)C^*) dt,
 \end{aligned}$$

where the expectation is with respect to the stationary measure.

5. Conclusion. The remaining problem for parameter identification for the TPBV process is to derive the practical algorithm for generating the optimal MLE. Noting that explicit form of smoother for the TPVB process has been derived in [4] in a computable form, it is possible to show the explicit form of necessary condition for the optimal MLE. Then, by using the similar technique presented by Aihara and Bagchi [10], from the derived necessary condition, we can construct a practical algorithm for generating the optimal MLE.

REFERENCES

- [1] M. B. ADAMS, A.S. WILLSKY, AND B. C. LEVY, *Linear estimation of boundary value stochastic process—Part I: The role and construction of complementary models*, IEEE Trans. Automat. Control, 29 (1984), pp. 803–811.
- [2] ———, *Linear estimation of boundary value stochastic process—Part II: 1-D smoothing problems*, IEEE Trans. Automat. Control, 29 (1984), pp. 811–822.
- [3] A. BAGCHI AND S. I. AIHARA, *Smoothing and identification for random fields*, in *Advances in Communication and Control Systems*, N. de Claris, ed., Optimization Software, Inc., New York, 1988, pp. 3–16.
- [4] A. BAGCHI AND H. WESTDIJK, *Smoothing and likelihood ratio for Gaussian boundary value processes*, IEEE Trans. Automat. Control, 34 (1989), pp. 954–962.
- [5] V. BORKAR AND A. BAGCHI, *Parameter estimation in continuous-time stochastic processes*, Stochastics, 8 (1982), pp. 193–212.
- [6] L. A. SHEPP, *Radon–Nikodym derivative of Gaussian measures*, Ann. Math. Statist., 37 (1966), pp. 321–354.
- [7] K. ITÔ, *Extension of stochastic integrals*, in *Proc. International Symposium on Stochastic Differential Equations*, K. Itô, ed., John Wiley, New York, 1978, pp. 95–109.
- [8] A. V. BALAKRISHNAN, *Stochastic Differential Systems*, Lecture Notes in Economics and Math. Systems, 84, Springer-Verlag, New York, 1973.
- [9] A. N. SHIRYAYEV, *Probability*, Springer-Verlag, New York, 1984.
- [10] S. I. AIHARA AND A. BAGCHI, *Parameter identification for stochastic diffusion equations with unknown boundary conditions*, Appl. Math. Optim., 15 (1988), pp. 15–36.

A METHOD OF CENTERS BASED ON BARRIER FUNCTIONS FOR SOLVING OPTIMAL CONTROL PROBLEMS WITH CONTINUUM STATE AND CONTROL CONSTRAINTS*

E. POLAK†, T. H. YANG†, AND D. Q. MAYNE‡

Abstract. This paper describes a method of centers based on barrier functions for solving optimal control problems with continuum inequality constraints on the state and control. The method decomposes the original problem into a sequence of easily solved optimal control problems with control constraints only. The method requires only approximate solution of these problems.

Key words. optimal control algorithms, method of centers, barrier functions, state space, control constraints

AMS(MOS) subject classifications. 49N40, 49M39, 49M30

1. Introduction. The difficulty of an optimal control problem is very much a function of the constraints. In the realm of optimal control problems, optimal control problems with control constraints and inequality state-space constraints rank close to the top in terms of difficulty or, alternatively, close to the bottom in terms of tractability. In this paper, we use ideas contained in recent work on phase I-phase II methods of centers [20], methods of centers based on barrier functions [7], [16], [22], and barrier function methods for semi-infinite minimax problems [21] to construct a reasonably promising optimal control algorithm for solving optimal control problems with both control and inequality state-space constraints. An important feature of this algorithm is that it decomposes the original problem into an infinite sequence of highly tractable optimal control problems with integral cost and control constraints only, each of which need only to be solved approximately.

To help establish the extent to which this paper advances the state of the art, we now discuss some of the earlier results in this area. We recall that free-time problems can always be transcribed into fixed-time problems by means of an augmentation of the dynamics (see [26]), and hence we need only discuss fixed-time problems. First (see [2]), unconstrained optimal control problems and optimal control problems with inequality endpoint constraints (both without control constraints), with smooth dynamics, can be formulated as

$$(1.1a) \quad \min_{u \in C} f^0(u)$$

and

$$(1.1b) \quad \min_{u \in C} \{f^0(u) \mid f^j(u) \leq 0, j = 1, 2, \dots, q\},$$

* Received by the editors May 7, 1990; accepted for publication (in revised form) November 12, 1991. The research reported herein was sponsored by Air Force Office of Scientific Research contract AFOSR-90-0068, National Science Foundation grant ECS-8816168, and the United Kingdom Science and Engineering Research Council.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720.

‡ Department of Electrical Engineering and Computer Sciences, University of California, Davis, California 95616 and Department of Electrical Engineering, Imperial College of Science and Technology, London, SW7 2BT, United Kingdom.

respectively, with all the functions $f^j(\cdot)$ continuously Frechet differentiable on the pre-Hilbert space $L_{\infty,2} \triangleq \{L_{\infty}^m[0,1], \|\cdot\|_2\}$, where $\|\cdot\|_2$ denotes the norm $L_2^m[0,1]$. These problems can be solved by exact analogues of the Armijo gradient method [1] (see, e.g., [8], [27]) and of the method of centers (see [17], [18]), respectively.

Simple optimal control problems with control constraints only assume the form

$$(1.2a) \quad \min_{u \in U} f^0(u),$$

where $U \triangleq \{u \in L_{\infty,2} | u(t) \in U, t \in [0,1]\}$, with U a compact subset of \mathbb{R}^m , and $f^0(\cdot)$ as above. For simple sets U , these problems can be solved by analogues of the slow-to-converge Frank-Wolfe algorithm [5], the faster Goldstein-Levitin-Polyak gradient projection method (see [3]), and a multiplier method [6], as well as by two algorithms that are optimal control specific: the strong variations algorithm in [23] and the relaxed-control steepest-descent algorithm in [27]. The addition of control constraints to (1.1b) results in a problem of the form

$$(1.2b) \quad \min_{u \in U} \{f^0(u) | f^j(u) \leq 0, j = 1, 2, \dots, q\},$$

which can be solved by an extension (see [12]) of the method of centers [17], [18]. However, the control constraints result in a considerable increase in difficulty in the search-direction-finding problem. The addition of equality constraints to (1.2b) can be handled by means of exact penalty functions (see, e.g., [12]–[14]).

The most difficult optimal control problems have both control and state-space constraints and can assume the following abstract form:

$$(1.3) \quad \min_{u \in U} \left\{ f^0(u) | f^j(u) \leq 0, j = 1, 2, \dots, q, \max_{t \in [0,1]} \phi^k(u, t) \leq 0, k = 1, 2, \dots, r \right\},$$

where the functions $f^j(\cdot)$ and $\phi^k(\cdot, \cdot)$ are all continuously differentiable. We recognize these problems as generalizations of finite-dimensional semi-infinite programming problems (see [19]). The presence of the control constraints generates a major obstacle because it precludes the efficient use of minimax theorems in the solution of extremely difficult search-direction-finding problems (see [19] for their use in semi-infinite optimization).

Not counting nonlinear programming algorithms on discretized versions of optimal control problems or algorithms whose convergence has not been established (see, e.g., [9], [10], [15]), there appear to be only two algorithms in the literature for the solution of problems of the form (1.3). They are both extensions of the method in [17], [18]; the one in [28] is based on the use of relaxed controls, whereas the one in [12] is not. Both of these algorithms postulate extremely difficult search-direction computations.

The algorithm we present in this paper is much simpler in structure than either of the algorithms [28], [29] or [11], [12]; furthermore, it is easily implemented using existing methods (such as in [3], [8]). In § 2 we present our algorithm in a simplified (conceptual) form. In §§ 3 and 4 we give full details of two alternative versions of our new algorithm and prove their convergence to feasible stationary points. Computational results are reported in § 5 and show that the algorithm performs satisfactorily.

2. A conceptual phase I–phase II method of centers. We consider optimal control problems defined in the pre-Hilbert space $L_{\infty,2} \triangleq \{L_{\infty}^m[0,1], \|\cdot\|_2\}$, consisting of elements in $L_{\infty}^m[0,1]$ but endowed with the $L_2[0,1]$ scalar product $\langle \cdot, \cdot \rangle_2$ and corresponding norm $\|\cdot\|_2$. The problems are normalized, fixed-time problems with control and

state-space constraints, of the form

$$(2.1a) \quad \mathbf{P}: \min \left\{ f^0(u) \mid f^j(u) \leq 0, j = 1, 2, \dots, q_1, \right. \\ \left. \max_{t \in [0, 1]} \phi^k(u, t) \leq 0, k = 1, 2, \dots, q_2, u \in G \right\},$$

where

$$(2.1b) \quad G \triangleq \{u \in L_{\infty, 2} \mid u(t) \in U, \forall t \in [0, 1]\}$$

with $U \subset \mathbb{R}^m$. The cost function $f^0: L_{\infty, 2} \rightarrow \mathbb{R}$ and the endpoint constraint functions $f^j: L_{\infty, 2} \rightarrow \mathbb{R}, j = 1, 2, \dots, q_1$ are defined by

$$(2.2a) \quad f^j(u) \triangleq g^j(x^u(1));$$

whereas the state-space constraint functions $\phi^k: L_{\infty, 2} \times [0, 1] \rightarrow \mathbb{R}, k = 1, 2, \dots, q_2$, are defined by

$$(2.2b) \quad \phi^k(u, t) \triangleq g^k(x^u(t)),$$

where the functions $g^j, g^k: \mathbb{R}^n \rightarrow \mathbb{R}$ and $x^u(\cdot)$ is the solution of the differential equation

$$(2.3a) \quad \dot{x}(t) = h(x(t), u(t)), \quad t \in [0, 1],$$

$$(2.3b) \quad x(0) = x_0,$$

where $h: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, and $x_0 \in \mathbb{R}^n$ is given.

We assume that the problem \mathbf{P} has a solution. In addition, we will require the following hypotheses, which ensure (see [2]) that (a) the solutions $x(\cdot)$ exist and are locally Lipschitz continuously differentiable, (b) the functions $f^j(\cdot)$ are locally Lipschitz continuously differentiable, and (c) the functions $\phi^j(\cdot, \cdot)$ are locally Lipschitz continuously differentiable in u and continuous in t .

Assumption 2.1. Let the following hold:

- (i) The set $U \subset \mathbb{R}^m$ is compact and convex;
- (ii) The functions $g^j, g^k: \mathbb{R}^n \rightarrow \mathbb{R}$ are locally Lipschitz continuously differentiable;
- (iii) The function $h: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is locally Lipschitz continuously differentiable;
- (iv) There is a constant $M < \infty$ such that $\|h(x, u)\| \leq M(1 + \|x\|)$ for all $(x, u) \in X \times U$, where X is a sufficiently large but bounded subset of \mathbb{R}^n .

For the purpose of convergence analysis, it is useful to introduce the relaxed-controls closure of the set G .

We recall that a Radon probability measure μ on the Borel sets of U (in (2.1b)) is a positive measure such that $\mu(U) = 1$. The set of Radon probability measures will be denoted by $\text{rpm}(U)$. A relaxed control σ is a measurable function $\sigma: [0, 1] \rightarrow \text{rpm}(U)$. We define the relaxed-controls closure of the set G (in (2.1b)) by

$$(2.4a) \quad \bar{G} \triangleq \{\sigma: [0, 1] \rightarrow \text{rpm}(U) \mid \sigma \text{ is measurable}\}.$$

We use the weak* topology on $L^1([0, 1], C(U))^*$ to topologize \bar{G} . Consequently, $\{\sigma_i\} \subset \bar{G}$ converges to $\sigma \in \bar{G}$ if and only if

$$(2.4b) \quad \lim_{i \rightarrow \infty} \int_0^1 \int_U \phi(t, u) \sigma_i(t)(du) dt = \int_0^1 \int_U \phi(t, u) \sigma(t)(du) dt$$

$$\forall \phi \in L^1([0, 1], C(U)),$$

where $L^1([0, 1], C(U))$ denotes the space of absolutely integrable functions that map the interval $[0, 1]$ into $C(U)$, the space of real-valued continuous functions defined on U .

In this topology, \bar{G} is sequentially compact. We recall that there is an injection of the ordinary controls into the relaxed controls: With each ordinary control $u \in G$, we associate a relaxed control $\sigma \in \bar{G}$ such that $\sigma(t)(S) = \delta_{u(t)}(S)$ for all measurable sets $S \subset U$, where $\delta_u(S) = 1$ if $u \in S$ and $\delta_u(S) = 0$ otherwise.

Relaxed controls give rise to relaxed dynamics as follows:

$$(2.4c) \quad \dot{x}(t) = \int_U h(x(t), u) \sigma(t)(du), \quad t \in [0, 1],$$

$$(2.4d) \quad x(0) = x_0,$$

whose solutions we will denote by $\bar{x}^\sigma(t)$. We extend this notation also to the functions in (2.2a), (2.2b); thus $\bar{f}^j(\sigma) = g^j(\bar{x}^\sigma(1))$ and $\bar{\phi}^k(\sigma, t) = g^k(\bar{x}^\sigma(t))$.

Our exposition will be simpler if we assume a single form for both the state-space and endpoint constraints. This requires that the functions $f^j(\cdot)$ be replaced by functions of the form $\max_{t \in [0, 1]} \phi^j(u, t)$ with $\phi^j(u, t) \triangleq g^j(x^u(1))$ for all $t \in [0, 1]$. Then, if we let $q = q_1 + q_2$ and replace the indices k in (2.2b) by $j = q_1 + k$, problem (2.1a) becomes

$$(2.5a) \quad \mathbf{P}: \min \left\{ f^0(u) \mid \max_{t \in [0, 1]} \phi^j(u, t) \leq 0, j = 1, 2, \dots, q, u \in G \right\},$$

or, in the even more compact form,

$$(2.5b) \quad \mathbf{P}: \min \{ f^0(u) \mid \psi^j(u) \leq 0, j = 1, 2, \dots, q, u \in G \},$$

where $\psi^j(u) \triangleq \max_{t \in [0, 1]} \phi^j(u, t)$.

We can also state the relaxed control version of (2.5b), as follows:

$$(2.5c) \quad \bar{\mathbf{P}}: \min \{ \bar{f}^0(\sigma) \mid \bar{\psi}^j(\sigma) \leq 0, j = 1, 2, \dots, q, \sigma \in \bar{G} \}.$$

Since we have assumed that \mathbf{P} has a solution, the *minimum values* for \mathbf{P} and $\bar{\mathbf{P}}$ are the same. However, it is conceivable that $\bar{\mathbf{P}}$ has solutions that do not have counterparts in G .

For any $u \in L_{\infty, 2}$, let $\psi(u) \triangleq \max_{j \in q} \psi^j(u)$, where $q \triangleq \{1, 2, \dots, q\}$, and let $\psi(u)_+ \triangleq \max \{0, \psi(u)\}$. The phase I-phase II methods of centers that we present in this paper are based on the use of the *unifying* function $F: L_{\infty, 2} \times L_{\infty, 2} \rightarrow \mathbb{R}$, defined by¹

$$(2.6) \quad F(u \mid u') = \max_{j \in q} \{ f^0(u) - f^0(u') - 2\psi(u')_+, \psi^j(u) - \psi(u')_+ \}.$$

The following result is obvious.

PROPOSITION 2.1. (i) For all $u \in L_{\infty, 2}$, $F(u \mid u) = 0$. (ii) Suppose that $\hat{u} \in G$ is a local optimizer for problem (2.5b). Then $F(u \mid \hat{u}) \geq 0$ for all $u \in G$ near \hat{u} , i.e., \hat{u} is a local minimizer for the problem $\min_{u \in G} F(u \mid \hat{u})$.

Phase I-phase II methods of centers are based on the following geometric notion: Given a point $u_i \in G$, its successor u_{i+1} is chosen to be a “center” of the set

$$(2.7) \quad V(u_i) \triangleq \{u \in G \mid F(u \mid u_i) \leq 0\}.$$

For our methods of centers to work, we must introduce the following commonly used “constraint qualification” type of hypothesis, which is easily interpreted in terms of Proposition 2.1.

¹ The scale factor 2 in the term $2\psi(u')_+$ in (2.6) can be replaced by any other scale factor $\gamma > 1$.

Assumption 2.2. For every $\sigma \in \bar{G}$ that is not a solution of $\bar{\mathbf{P}}$, there exists a $u \in G$ such that $F(u|\sigma) < 0$.

The methods differ by the manner in which they define a “center.” The simplest, but not practical, definition of the “center” u_{i+1} is

$$(2.8) \quad u_{i+1} = \operatorname{argmin}_{u \in G} F(u|u_i).$$

Solving (2.8) for u_{i+1} is hardly easier than solving the original problem (2.5b). Hence, we will now introduce a much more tractable definition of a “center” based on the parametrized *barrier function* $p_\alpha : L_{\infty,2} \times L_{\infty,2} \rightarrow \mathbb{R}$ for the above sets $V(u')$, defined by, for $\alpha > 0$,

$$(2.9) \quad p_\alpha(u|u') \triangleq \frac{1}{\alpha + 2\psi(u')_+ + f^0(u') - f^0(u)} + \sum_{j=1}^q \int_0^1 \frac{1}{\alpha + \psi(u')_+ - \phi^j(u, t)} dt.$$

Because for $\alpha > 0$, $p_\alpha(u'|u') < \infty$, the parameter α makes it possible to use the point u' for initializing a descent method in solving $\min_{u \in V(u')} p_\alpha(u|u')$.

We begin by establishing that as $\alpha \rightarrow 0$, $p_\alpha(\cdot|u')$ becomes a barrier function for the set $V(u')$.

LEMMA 2.1. *There exists a constant $L > 0$, such that, for all $u' \in G$, $u \in V(u')$, and $\alpha > 0$,*

$$(2.10) \quad p_\alpha(u|u') \geq \frac{1}{\alpha + 2\psi(u')_+ + f^0(u') - f^0(u)} + \frac{1}{L} \log \left(1 + \frac{L}{(\alpha + \psi(u')_+ - \psi(u))} \right).$$

Proof. It follows from our assumptions that there exists a constant $L < \infty$, such that each $\phi^j(u, \cdot)$ is uniformly Lipschitz in t on $[0, 1]$ with the same Lipschitz constant L for all $u \in G$. Without loss of generality, we may assume that $L \geq \psi(u')_+ - \psi(u)$ for all $u, u' \in G$. Now, given $u' \in G$ and $u \in V(u')$, let $k \in \bar{q}$ and $\hat{t} \in [0, 1]$ be such that $\phi^k(u, \hat{t}) = \psi(u)$, and let $t \in [0, 1]$ be arbitrary. Then we have that

$$(2.11) \quad \phi^k(u, t) \geq \phi^k(u, \hat{t}) - L|t - \hat{t}| = \psi(u) - L|t - \hat{t}|.$$

Consequently, since $\psi(u')_+ - \psi(u) \geq 0$,

$$(2.12) \quad \begin{aligned} p_\alpha(u|u') &\geq \frac{1}{\alpha + 2\psi(u')_+ + f^0(u') - f^0(u)} + \int_0^1 \frac{1}{(\alpha + \psi(u')_+ - \phi^k(u, t))} dt \\ &\geq \frac{1}{\alpha + 2\psi(u')_+ + f^0(u') - f^0(u)} + \int_0^1 \frac{1}{(\alpha + \psi(u')_+ - \psi(u) + L|t - \hat{t}|)} dt \\ &\geq \frac{1}{\alpha + 2\psi(u')_+ + f^0(u') - f^0(u)} \\ &\quad + \frac{1}{L} \log \left(\frac{(\alpha + \psi(u')_+ - \psi(u) + L\hat{t})(\alpha + \psi(u')_+ - \psi(u) + L(1 - \hat{t}))}{(\alpha + \psi(u')_+ - \psi(u))^2} \right) \\ &\geq \frac{1}{\alpha + 2\psi(u')_+ + f^0(u') - f^0(u)} + \frac{1}{L} \log \left(1 + \frac{L}{(\alpha + \psi(u')_+ - \psi(u))} \right), \end{aligned}$$

where the last line is obtained by minimizing the preceding line with respect to \hat{t} , which happens when $\hat{t} = 0$ or 1 . \square

It follows by inspection of (2.10) that when $u \rightarrow u'$ with $u \in V(u')$, then $p_0(u|u') \rightarrow \infty$, i.e., that $p_0(\cdot, u')$ is indeed a barrier function for $V(u')$.

Now consider the following conceptual algorithm for solving the problem **P** (2.5a).

Algorithm 2.1.

Data: $u_0 \in G$ and a sequence $\{\alpha_k\}_{k=0}^\infty$ such that $\alpha_k > 0$ for all $k \in \mathbb{N}$ and $\alpha_k \downarrow 0$ as $k \rightarrow \infty$.

Step 0: Set $i = 0$ and $k = 0$.

Step 1: Compute

$$(2.13) \quad u_{i+1} \in A(u_i) \triangleq \operatorname{argmin}_{u \in V(u_i)} p_{\alpha_k}(u|u_i).$$

Step 2: If $F(u_{i+1}|u_i) = 0$, replace k by $k+1$ and go to Step 1.

Step 3: Replace i by $i+1$ and k by $k+1$, and go to Step 1.

Note that (2.13) defines u_{i+1} as a solution of the simple optimal control problem, shown below:

$$(2.14) \quad \min_{u \in G} \left\{ \frac{1}{\alpha_{k_i} + 2\psi(u_i)_+ + g^0(x^{u_i}(1)) - g^0(x^u(1))} + \sum_{j=1}^q \int_0^1 \frac{1}{\alpha_{k_i} + \psi(u_i)_+ - g^j(x^u(t))} dt \right\},$$

where, as before, $x^u(t)$ is the solution of (2.4a), (2.4b). This problem has only control constraints; its cost is the endpoint-plus-integral form. Barring possible ill conditioning, such problems are tractable by algorithms such as the Goldstein-Levitin-Polyak gradient projection method (see [3]) or by the algorithm described in [2]. However, these methods require an *infinite* number of iterations to produce an accumulation point that can only be shown to be stationary (not necessarily optimal); hence Algorithm 2.1 is only conceptual.

THEOREM 2.1. (i) *Suppose that Algorithm 2.1 jams up in the loop between Steps 1 and 2 at the control u_{i_0} . Then u_{i_0} is a solution of **P**.* (ii) *Suppose that $\{u_i\}_{i=0}^\infty$ is a sequence of controls constructed by Algorithm 2.1. If this sequence has an accumulation point $\hat{u} \in G$, then \hat{u} is a solution of **P**.*

Proof. (i) For contradiction, suppose that u_{i_0} is not a solution of **P**. Since the algorithm is cycling in the loop between Steps 1 and 2, $k \rightarrow \infty$. Let $\hat{u}_k \triangleq \operatorname{argmin}_{u \in V(u_{i_0})} p_{\alpha_k}(u|u_{i_0})$. Then, since $\alpha_k \rightarrow 0$ as $k \rightarrow \infty$ and $F(\hat{u}_k|u_{i_0}) = 0$ for all k , we conclude that $p_{\alpha_k}(\hat{u}_k|u_{i_0}) \rightarrow \infty$ as $k \rightarrow \infty$. By Assumption 2.2, however, since u_{i_0} is not a solution of **P**, there exists a $\hat{u} \in V(u_{i_0})$ such that $F(\hat{u}|u_{i_0}) < 0$, which implies that $p_0(\hat{u}|u_{i_0}) < \infty$. Since $p_{\alpha_k}(\hat{u}|u_{i_0})$ is continuous in α , there exists a k_0 such that $p_{\alpha_k}(\hat{u}|u_{i_0}) \leq 2p_0(\hat{u}|u_{i_0})$ for all $k \geq k_0$. Because $p_{\alpha_k}(\hat{u}_k|u_{i_0}) \rightarrow \infty$, however, there exists a k_1 such that $p_{\alpha_k}(\hat{u}_k|u_{i_0}) > p_{\alpha_k}(\hat{u}|u_{i_0})$ for all $k \geq k_1 \geq k_0$, which contradicts the fact that \hat{u}_k is a minimizer.

(ii) Suppose that there exists an infinite subset $K \subset \mathbb{N}$ and a $\hat{u} \in G$ such that the subsequence $\{u_i\}_{i \in K}$ converges to \hat{u} ; we write this as $u_i \xrightarrow{K} \hat{u} \in G$ as $i \rightarrow \infty$. For contradiction, suppose that \hat{u} is not a solution of **P**.

Case 1. There exists an i_0 such that $\psi(u_{i_0}) \leq 0$, so that $\psi(u_{i_0})_+ = 0$. Since $F(u_{i_0+1}|u_{i_0}) < 0$, we have that

$$(2.15a) \quad f^0(u_{i_0+1}) - f^0(u_{i_0}) < 0,$$

$$(2.15b) \quad \psi(u_{i_0+1}) < 0.$$

It now follows by induction that, for all $i \geq i_0$,

$$(2.16a) \quad f^0(u_{i+1}) - f^0(u_i) < 0,$$

$$(2.16b) \quad \psi(u_i) < 0.$$

Hence, since the sequence $\{f^0(u_i)\}_{i=i_0}^\infty$ is monotone decreasing, and since $f^0(u_i) \xrightarrow{K} f^0(\hat{u})$ as $i \rightarrow \infty$, by continuity, it follows that $f^0(u_i) \rightarrow f^0(\hat{u})$ as $i \rightarrow \infty$; therefore, since

$$(2.17a) \quad p_{\alpha_{k_i}}(u_{i+1}|u_i) > \frac{1}{\alpha_{k_i} + f^0(u_i) - f^0(u_{i+1})},$$

and $\alpha_{k_i} \rightarrow 0$ as $i \rightarrow \infty$, we have that

$$(2.17b) \quad p_{\alpha_{k_i}}(u_{i+1}|u_i) \rightarrow \infty \quad \text{as } i \rightarrow \infty.$$

However, since \hat{u} is not a solution to **P**, by Assumption 2.2 there exists a $u^* \in V(\hat{u})$ such that $F(u^*|\hat{u}) < 0$ and therefore that $p_0(u^*|\hat{u}) < \infty$. Since, by construction, $V(\hat{u}) \subset V(u_i)$ for all $i \geq i_0$, it follows that

$$(2.18) \quad p_{\alpha_{k_i}}(u_{i+1}|u_i) \leq p_{\alpha_{k_i}}(u^*|u_i) \quad \forall i \geq i_0.$$

Furthermore, by continuity of $p_\alpha(u^*|u)$ in (α, u) , $p_{\alpha_{k_i}}(u^*|u_i) \xrightarrow{K} p_0(u^*|\hat{u}) < \infty$; hence, there exists an $i_1 \geq i_0$ such that, for all $i \in K$, $i \geq i_1$,

$$(2.19) \quad p_{\alpha_{k_i}}(u_{i+1}|u_i) \leq 2p_0(u^*|\hat{u}),$$

which is a contradiction of (2.17b).

Case 2. Suppose that $\psi(u_i) > 0$ for all $i \in \mathbb{N}$. Then, since $p_{\alpha_{k_i}}(u_{i+1}|u_i)$ for all $i \in \mathbb{N}$, it follows from (2.10) that $\psi(u_{i+1}) < \psi(u_i)$ for all $i \in \mathbb{N}$ and hence, since by continuity $\psi(u_i) \xrightarrow{K} \psi(\hat{u})$ as $i \rightarrow \infty$, that $\psi(u_i) \rightarrow \psi(\hat{u})$ as $i \rightarrow \infty$. Hence, since by (2.10)

$$(2.20a) \quad p_{\alpha_{k_i}}(u_{i+1}|u_i) > \frac{1}{L} \log \left(1 + \frac{L}{(\alpha_{k_i} + \psi(u_i)_+ - \psi(u_{i+1}))} \right)$$

and $\alpha_{k_i} \rightarrow 0$ as $i \rightarrow \infty$, we have that

$$(2.20b) \quad p_{\alpha_{k_i}}(u_{i+1}|u_i) \rightarrow \infty \quad \text{as } i \rightarrow \infty.$$

However, by Assumption 2.2, since \hat{u} is not a solution of **P**, there exists a $u^* \in V(\hat{u})$ such that $F(u^*|\hat{u}) < 0$, and hence $p_0(u^*|\hat{u}) < \infty$. By continuity of $F(u^*|\cdot)$, it now follows that there exists an i_2 such that $F(u^*|u_i) < 0$ for all $i \in K$, $i \geq i_2$, and hence we see that $u^* \in V(u_i)$ for all $i \in K$, $i \geq i_2$. Hence, again by continuity, there exists an $i_3 \geq i_2$ such that

$$(2.21) \quad p_{\alpha_{k_i}}(u_{i+1}|u_i) \leq p_{\alpha_{k_i}}(u^*|u_i) \leq 2p_0(u^*|\hat{u}) < \infty,$$

which contradicts (2.20b). \square

Since the set G is not compact, it is entirely possible that a sequence $\{u_i\}_{i=0}^\infty$ constructed by Algorithm 2.1 has no accumulation points in G . In that case, Theorem 2.1 is vacuous. However, the sequence $\{u_i\}_{i=0}^\infty$ must have accumulation points, in the sense of control measures, in the sequentially compact set \bar{G} . The following result follows directly from the arguments used to prove Theorem 2.1.

COROLLARY 2.1. *Suppose that $\{u_i\}_{i=0}^\infty$ is a sequence of controls constructed by Algorithm 2.1. If this sequence has an accumulation point $\hat{\sigma} \in \bar{G}$, then $\hat{\sigma}$ is a solution of $\bar{\mathbf{P}}$.*

3. An implementable phase I–phase II method of centers. The main objection to Algorithm 2.1 is that the update operation in (2.13) is not implementable. We will now develop an implementable algorithm that replaces (2.13) by an approximate stationarity condition and incorporates a feature that enables us to use the point u_i as a starting point in computing u_{i+1} by an algorithm such as Algorithm 5.14 in [2].

First, referring to [2], we see that the Frechet differentials (of the functions $f^0(\cdot)$, $\phi^k(\cdot, t)$ in (2.5a)) $df^0(u; u' - u)$, $d\phi^k(u, t; u' - u)$ exist and can be expressed in terms of scalar products with gradients $\nabla f^0(u)$ and $\nabla_u \phi^j(u, t)$, which are in $L_{\infty,2}$, i.e., $df^0(u; u' - u) = \langle \nabla f^0(u), u' - u \rangle_2$ and $d\phi^j(u, t; u' - u) = \langle \nabla_u \phi^j(u, t), u' - u \rangle_2$. Neither our proofs nor our algorithm require formulae for these gradients.

Next, we define an *optimality function* $\theta: G \rightarrow \mathbb{R}$ for the problem **P** (2.5a), which is a first-order convex approximation to the function $F(u|u)$, by

$$\begin{aligned} \theta(u) \triangleq \min_{u' \in G} \max \{ \frac{1}{2} \|u' - u\|_2^2 + \{-2\psi(u)_+ + df^0(u; u' - u), \\ (3.1) \quad \phi^j(u, t_j) - \psi(u)_+ + \langle \nabla_u \phi^j(u, t_j), u' - u \rangle_2, \\ t_j \in [0, 1], j \in q \} \}. \end{aligned}$$

At one point in our convergence proof, we will need to bring in the relaxed-controls topology. Hence, we need the relaxed-controls extension of $\theta(u)$. For this purpose, it is useful to recall that alternative formulae for $\langle \nabla f^0(u), u' - u \rangle_2$ and $\langle \nabla \phi^j(u, t), u' - u \rangle_2$ are given by

$$(3.2) \quad \langle \nabla f^0(u), u' - u \rangle_2 = \langle \nabla g^0(x^u(1)), \delta x^{u',u}(1) \rangle,$$

$$(3.3) \quad \langle \nabla \phi^j(u, t), u' - u \rangle_2 = \langle \nabla g^j(x^u(t)), \delta x^{u',u}(t) \rangle,$$

where $\delta x^{u',u}(t)$ is the solution of the first variational equation

$$\begin{aligned} (3.4) \quad \delta \dot{x}(t) = \left[\frac{\partial h(x^u(t), u(t))}{\partial x} \right] \delta x(t) \\ + \left[\frac{\partial h(x^u(t), u(t))}{\partial u} \right] [u'(t) - u(t)], \quad \delta x(0) = 0. \end{aligned}$$

As in [30] and [2], given a relaxed control $\sigma \in \bar{G}$ with corresponding solution $\bar{x}^\sigma(\cdot)$ of (2.4c), (2.4d) and a function $s: U \rightarrow L_{2,\infty}$, we define $\delta \bar{x}^{\sigma,s}(\cdot)$ to be the solution of

$$\begin{aligned} (3.5a) \quad \delta \dot{x}(t) = \int_U \left[\frac{\partial h(\bar{x}^\sigma(t), u)}{\partial x} \right] \sigma(t)(du) \delta x(t) \\ + \int_U \left[\frac{\partial h(x^\sigma(t), u)}{\partial u} \right] s(u)(t) \sigma(t)(du), \end{aligned}$$

$$(3.5b) \quad \delta x(0) = 0.$$

Next, we say that a *search direction function* $s: U \rightarrow L_{2,\infty}$ is *admissible* if, for all $u \in U$, $u + s(u)(t) \in U$ for almost all $t \in [0, 1]$. We define S to be the set of all admissible search direction functions. These definitions allow the relaxed-controls extension $\bar{\theta}: \bar{G} \rightarrow \mathbb{R}$ of $\theta(\cdot)$ to be defined by

$$\begin{aligned} \bar{\theta}(\sigma) \triangleq \min_{s \in S} \max \{ \frac{1}{2} \|s\|_2^2 + \{-2\bar{\psi}(\sigma)_+ + \langle \nabla g^0(\bar{x}^\sigma(1)), \delta \bar{x}^{\sigma,s}(1) \rangle, \\ (3.6) \quad \bar{\phi}^j(\sigma, t_j) - \bar{\psi}(\sigma)_+ + \langle \nabla g^j(\bar{x}^\sigma(t_j)), \delta \bar{x}^{\sigma,s}(t_j) \rangle, \\ t_j \in [0, 1], j \in q \} \}. \end{aligned}$$

The following result can be found in [2].

THEOREM 3.1. *We have the following conditions:*

- (i) *The optimality functions $\theta(\cdot)$ and $\bar{\theta}(\cdot)$ are well defined and continuous;*
- (ii) *If $\sigma \in \bar{G}$ corresponds to the ordinary control $u \in G$, then $\bar{\theta}(\sigma) = \theta(u)$;*
- (iii) *If $\hat{u} \in G$ is an optimal solution to the problem P , then $\theta(\hat{u}) = 0$;*
- (iv) *If $\hat{\sigma} \in \bar{G}$ is an optimal solution to the problem \bar{P} , then $\bar{\theta}(\hat{\sigma}) = 0$.*

In our proofs, we find it convenient to use an alternative formula for $\theta(u)$. Let the set of Radon probability measures on the interval $[0, 1]$ be denoted by $\text{rpm}([0, 1])$, let V denote the set of measurable functions $\nu: [0, 1] \rightarrow \text{rpm}([0, 1])$, and let

$$(3.7a) \quad W \triangleq \left\{ w = (w^0, w^1, \dots, w^q) \in \mathbb{R}^{q+1} \mid \sum_{i=0}^q w^i = 1, w^i \geq 0, i = 0, 1, \dots, q \right\}.$$

Finally, with $\text{rfm}([0, 1])$ denoting the space of Radon finite measures, let Σ be the set of measurable functions $\mu: [0, 1] \rightarrow [\text{rfm}([0, 1])]^{q+1}$ defined by

$$(3.7b) \quad \Sigma \triangleq \{ \mu \in [\text{rfm}([0, 1])]^{q+1} \mid \mu^j = w^j \nu, j = 0, 1, \dots, q, w \in W, \nu \in V \}.$$

Then it is obvious that

$$(3.7c) \quad \theta(u) \triangleq \min_{u' \in G} \max_{\mu \in \Sigma} \left\{ \frac{1}{2} \|u' - u\|_2^2 + \left\{ \int_{[0,1]} [-2\psi(u)_+ + df^0(u; u' - u)] \mu^0(t) (dt) \right. \right. \\ \left. \left. + \sum_{j=1}^q \int_{[0,1]} [\phi^j(u, t) - \psi(u)_+] \mu^j(t) (dt) \right. \right. \\ \left. \left. + \sum_{j=1}^q \left\langle \int_{[0,1]} \nabla_u \phi^j(u, t) \mu^j(t) (dt), u' - u \right\rangle_2 \right\} \right\}.$$

We will require the following assumption, which is usually required for methods of centers and feasible directions.

Assumption 3.1. (i) The \bar{G} -closure of the set $\{u \in G \mid \psi(u) \leq 0\}$ is equal to the \bar{G} -closure of its interior. (ii) For all $u \in G$ such that $\psi(u) > 0$, $\theta(u) < 0$.

Next, referring to definition (2.9), we conclude that

$$(3.8a) \quad \nabla_u p_\alpha(u \mid u') = \frac{\nabla f^0(u)}{[\alpha + 2\psi(u')_+ + f^0(u') - f^0(u)]^2} \\ + \sum_{j=1}^q \int_0^1 \frac{\nabla_u \phi^j(u, t)}{[\alpha + \psi(u')_+ - \phi^j(u, t)]^2} dt.$$

To evaluate $\nabla_u p_\alpha(u \mid u')$, we do not use the cumbersome formula (3.8a); rather, we use the following computationally efficient formula:

$$(3.8b) \quad \nabla_u p_\alpha(u \mid u')(t) = \left[\frac{\partial h(x^u(t), u(t))}{\partial u} \right]^T \lambda^{u'u}(t),$$

where $\lambda^{u'u}(t)$ is the solution of the following adjoint system:

$$(3.8c) \quad \dot{\lambda}(t) = - \left[\frac{\partial h(x^u(t), u(t))}{\partial x} \right]^T \lambda(t) \\ - \sum_{j=1}^q \frac{\nabla_x g^j(x^u(t))}{[\alpha + \psi(u')_+ - g^j(x^u(t))]^2}, \quad t \in [0, 1],$$

$$(3.8d) \quad \lambda(1) = \frac{1}{[\alpha + 2\psi(u')_+ + g^0(x^{u'}(1)) - g^0(x^u(1))]^2} \nabla g^0(x^u(1)).$$

Algorithm 2.1 now gives rise to the following implementable algorithm.

Algorithm 3.1.

Data: $u_0 \in G$, $\varepsilon > 0$, and $\{\alpha_k\}_{k=0}^\infty$ such that $\alpha_k > 0$ for all $k \in \mathbb{N}$ and $\alpha_k \downarrow 0$ as $k \rightarrow \infty$.

Step 0: Set $i = 0$ and $k = 0$.

Step 1: Use any descent algorithm to generate a $u_{i+1} \in V(u_i)$ such that

$$(3.9) \quad 0 \geq \min_{u \in G} \langle \nabla_u p_{\alpha_k}(u_{i+1} | u_i), u - u_{i+1} \rangle_2 + \frac{1}{2} \|u - u_{i+1}\|^2 \geq -\varepsilon.$$

Step 2: If $F(u_{i+1} | u_i) = 0$, replace k by $k + 1$ and go to Step 1.

Step 3: Replace i by $i + 1$ and k by $k + 1$, and go to Step 1.

The proof of convergence of Algorithm 3.1 depends on the following two lemmas.

LEMMA 3.1. *Suppose that Assumption 2.1 holds, that $\gamma > 0$, that $u' \in G$, and that $u \in V(u')$. For $j \in q$, let $T_\gamma^j(u) \subset [0, 1]$ be defined by*

$$(3.10a) \quad T_\gamma^j(u) = \{t \in [0, 1] \mid \phi^j(u, t) \geq \psi(u) - \gamma\}.$$

Then, for each $t \notin T_\gamma^j(u)$, with $t \in [0, 1]$ and $j \in q$,

$$(3.10b) \quad \frac{1}{(\psi(u')_+ - \phi^j(u, t))} \leq \frac{1}{\gamma}.$$

Proof. Because $t \notin T_\gamma^j(u)$, $\phi^j(u, t) < \psi(u) - \gamma$. Since $u \in V(u')$, we obtain that $\psi(u')_+ \geq \psi(u)$. Hence,

$$(3.11) \quad \psi(u')_+ - \phi^j(u, t) \geq \psi(u) - \phi^j(u, t) \geq \gamma,$$

and the desired inequality follows. \square

LEMMA 3.2. *Suppose that Assumption 2.1 holds. Then, for any $M_\alpha < \infty$, there exists a constant $\delta > 0$ such that, for any $0 < \alpha \leq M_\alpha$, for all $u' \in G$, and for all $u \in V(u')$,*

$$(3.12) \quad (\alpha + \psi(u')_+ - \psi(u)) \sum_{j \in q} \int_{[0, 1]} \frac{1}{(\alpha + \psi(u')_+ - \phi^j(u, t))^2} dt \geq \delta > 0.$$

Proof. Since G is bounded, it follows from Assumption 2.1 that there exists a Lipschitz constant $L < \infty$ such that each $\phi^j(u, \cdot)$ is uniformly Lipschitz in t on $[0, 1]$ for all $u \in G$. Without loss of generality, we may assume that $L \geq \psi(u')_+ - \psi(u)$ for all $u, u' \in G$. Let $j \in q$ be such that $T_0^j(u)$ is nonempty, let $t_u \in T_0^j(u)$ be given and let $t \in [0, 1]$. Then we have that

$$(3.13a) \quad \phi^j(u, t) \geq \phi^j(u, t_u) - L|t - t_u| = \psi(u) - L|t - t_u|.$$

Now suppose that $0 < \gamma \leq L + M_\alpha$. Then $\{t \in [0, 1] \mid |t - t_u| \leq \gamma/(L + M_\alpha)\} \subset T_\gamma^j(u)$, and hence $m(T_\gamma^j(u)) \geq \gamma/(L + M_\alpha)$, where $m(\cdot)$ denotes the Lebesgue measure on \mathbb{R} . Hence, we conclude that

$$(3.13b) \quad \begin{aligned} \sum_{j \in q} \int_{[0, 1]} \frac{\alpha + \psi(u')_+ - \psi(u)}{(\alpha + \psi(u')_+ - \phi^j(u, t))^2} dt &\geq \int_{T_\gamma^j(u)} \frac{\alpha + \psi(u')_+ - \psi(u)}{(\alpha + \psi(u')_+ - \phi^j(u, t))^2} dt \\ &\geq \frac{\gamma}{(L + M_\alpha)} \frac{\alpha + \psi(u')_+ - \psi(u)}{(\alpha + \psi(u')_+ - \psi(u) + \gamma)^2}. \end{aligned}$$

Setting $\gamma = \alpha + \psi(u')_+ - \psi(u) \leq M_\alpha + L$ and $\delta = \frac{1}{4}(L + M_\alpha)$, we obtain the desired result. \square

THEOREM 3.2. (i) *Suppose that Algorithm 3.1 jams up in the loop between Steps 1 and 2 at the control u_{i_0} . Then $\psi(u_{i_0}) \leq 0$ and $\theta(u_{i_0}) = 0$.* (ii) *Suppose that $\{u_i\}_{i=0}^\infty$ is a sequence of controls constructed by Algorithm 3.1. If this sequence has an accumulation point $\hat{u} \in G$, then $\psi(\hat{u}) \leq 0$ and $\theta(\hat{u}) = 0$.*

Proof. (i) The proof of this part is essentially the same as for (i) of Theorem 2.1 and hence is omitted. (ii) Since u_{i+1} is constructed from u_i by a descent method, it follows from $F(u_{i+1}|u_i) < 0$ that, if $\psi(u_i) \leq 0$, then $\psi(u_{i+1}) \leq 0$ also. Hence, our proof breaks down into the examination of two cases.

Case 1. Suppose that $u_i \xrightarrow{K} \hat{u} \in G$ as $i \rightarrow \infty$ and that there exists an i_0 such that $\psi(u_i) \leq 0$ for all $i \geq i_0$. Then, by construction, the sequence $\{f^0(u_i)\}_{i=i_0}^\infty$ is monotone decreasing, and, since it is bounded, it must converge to $f(\hat{u})$. The fact that $\psi(\hat{u}) \leq 0$ follows directly from the continuity of $\psi(\cdot)$.

Now, for $i \geq i_0$, $j \in q$, and $t \in [0, 1]$, let

$$(3.14a) \quad \rho_i^j(t) \triangleq \frac{[\alpha_{k_i} + f^0(u_{i-1}) - f^0(u_i)]^2}{(\alpha_{k_i} - \phi^j(u_i, t))^2}.$$

Finally, let

$$(3.14b) \quad \nu_i \triangleq 1 + \sum_{j=1}^q \int_0^1 \rho_i^j(t) dt.$$

It now follows from (3.8a) and (3.9) that, for all $i \geq i_0$,

$$(3.15a) \quad \begin{aligned} 0 \geq \Theta'(u_i) &\triangleq \min_{u \in G} \frac{1}{\nu_i} \left\{ df^0(u_i; u - u_i) + \sum_{j=1}^q \int_0^1 \rho_i^j(t) d\phi^j(u_i, t; u - u_i) + \frac{1}{2} \|u - u_i\|_2^2 \right\} \\ &\geq -\varepsilon_i \triangleq -\varepsilon \frac{[\alpha_{k_i} + f^0(u_{i-1}) - f^0(u_i)]^2}{\nu_i}. \end{aligned}$$

Since $\nu_i \geq 1$ for all $i \geq i_0$ and since $f^0(u_i) \rightarrow f^0(\hat{u})$ and $\alpha_{k_i} \rightarrow 0$ as $i \rightarrow \infty$, it follows that $\varepsilon_i \rightarrow 0$ as $i \rightarrow \infty$. Next, it follows from (3.7c) that, for all $i \geq i_0$,

$$(3.15b) \quad \begin{aligned} 0 \geq \theta(u_i) &\geq \Theta'(u_i) + \frac{1}{\nu_i} \sum_{j=1}^q \int_0^1 \rho_i^j(t) \phi^j(u_i, t) dt \\ &\geq -\varepsilon_i + \frac{1}{\nu_i} \sum_{j=1}^q \int_0^1 \rho_i^j(t) \phi^j(u_i, t) dt. \end{aligned}$$

Now, for any $\gamma > 0$, let $I_\gamma^j(u) \subset [0, 1]$ be defined by

$$(3.16a) \quad I_\gamma^j(u) \triangleq \{t \in [0, 1] \mid \phi^j(u, t) \geq -\gamma/2\}.$$

It now follows from (3.14a, b) and the relation

$$0 \geq \frac{\phi^j(u_i, t)}{(\alpha_{k_i} - \phi^j(u_i, t))^2} \geq \frac{\phi^j(u_i, t)}{(-\phi^j(u_i, t))^2}$$

that, for any $j \in q$, and any $\gamma > 0$,

$$(3.16b) \quad \begin{aligned} 0 &\geq \frac{1}{\nu_i} \int_0^1 \rho_i^j(t) \phi^j(u_i, t) dt \\ &= \frac{1}{\nu_i} \int_{t \in I_\gamma^j(u_i)} \rho_i^j(t) \phi^j(u_i, t) dt + \frac{1}{\nu_i} \int_{t \in I_\gamma^j(u_i)^c} \rho_i^j(t) \phi^j(u_i, t) dt \\ &\geq -\frac{\gamma}{2} + \frac{[\alpha_{k_i} + f^0(u_{i-1}) - f^0(u_i)]^2}{\nu_i} \int_{t \in I_\gamma^j(u_i)^c} \frac{\phi^j(u_i, t)}{(\alpha_{k_i} - \phi^j(u_i, t))^2} dt \\ &\geq -\frac{\gamma}{2} + \frac{[\alpha_{k_i} + f^0(u_{i-1}) - f^0(u_i)]^2}{\nu_i} \left\{ \int_{t \in I_\gamma^j(u_i)^c} \frac{1}{\phi^j(u_i, t)} dt \right\} \\ &\geq -\frac{\gamma}{2} - \left(\frac{[\alpha_{k_i} + f^0(u_{i-1}) - f^0(u_i)]^2}{\nu_i} \right) \frac{2}{\gamma}. \end{aligned}$$

Since $[\alpha_{k_i} + f^0(u_{i-1}) - f^0(u_i)]^2 / \nu_i \rightarrow 0$ as $i \rightarrow \infty$, (3.16b) holds for any $\gamma > 0$ and, by Theorem 3.1, $\theta(\cdot)$ is continuous, it now follows that $\theta(\hat{u}) = 0$.

Case 2. We now suppose that $\psi(u_i) > 0$ for all $i \in \mathbb{N}$. Then we must have that the sequence $\{\psi(u_i)\}_{i=0}^\infty$ is monotonically decreasing, and hence it must converge to $\psi(\hat{u})$. We note that $\psi(u_i) = \psi(u_i)_+$ for all $i \in \mathbb{N}$. In this case, we define

$$(3.17a) \quad \rho_i^0(t) \triangleq \frac{\alpha_{k_i} + \psi(u_{i-1}) - \psi(u_i)}{(\alpha_{k_i} + 2\psi(u_{i-1}) + f^0(u_{i-1}) - f^0(u_i))^2},$$

$$(3.17b) \quad \rho_i^j(t) \triangleq \frac{\alpha_{k_i} + \psi(u_{i-1}) - \psi(u_i)}{(\alpha_{k_i} + \psi(u_{i-1}) - \phi^j(u_i, t))^2}, \quad j = 1, 2, \dots, q.$$

It now follows from Lemma 3.2 that

$$(3.18) \quad \nu_i \triangleq \sum_{j=0}^q \int_0^1 \rho_i^j(t) dt \geq \delta > 0$$

and from (3.9) that

$$(3.19) \quad \begin{aligned} 0 \geq \Theta''(u_i) &\triangleq \min_{u \in G} \frac{1}{\nu_i} \left\{ \int_0^1 \rho_i^0(t) df^0(u_i; u - u_i) dt \right. \\ &\quad \left. + \sum_{j=1}^q \int_0^1 \rho_i^j(t) d\phi^j(u_i, t; u - u_i) dt + \frac{1}{2} \|u - u_i\|_2^2 \right\} \\ &\geq -\varepsilon_i \triangleq -\varepsilon \frac{[\alpha_{k_i} + \psi(u_{i-1}) - \psi(u_i)]}{\nu_i}. \end{aligned}$$

Clearly, $\varepsilon_i \rightarrow 0$ as $i \rightarrow \infty$. Next, by the same argument as in (3.15b), we obtain that

$$(3.20) \quad \begin{aligned} 0 \geq \theta(u_i) &\geq \Theta''(u_i) + \frac{1}{\nu_i} \left\{ \int_0^1 -2\rho_i^0(t)\psi(u_{i-1}) dt \right. \\ &\quad \left. + \sum_{j=1}^q \int_0^1 \rho_i^j(t)[\phi^j(u_i, t) - \psi(u_{i-1})] dt \right\} \\ &\geq -\varepsilon_i + \frac{1}{\nu_i} \left\{ \int_0^1 -2\rho_i^0(t)\psi(u_{i-1}) dt \right. \\ &\quad \left. + \sum_{j=1}^q \int_0^1 \rho_i^j(t)[\phi^j(u_i, t) - \psi(u_{i-1})] dt \right\}. \end{aligned}$$

First, it follows from (3.18), the relation

$$0 \geq \frac{\phi^j(u_i, t) - \psi(u_{i-1})}{(\alpha_{k_i} + \psi(u_{i-1}) - \phi^j(u_i, t))^2} \geq \frac{1}{\phi^j(u_i, t) - \psi(u_{i-1})},$$

and Lemma 3.1 that, for all $j \in \underline{q}$ and for any $\gamma > 0$,

$$\begin{aligned} 0 &\geq \frac{1}{\nu_i} \int_0^1 \rho_i^j(t)[\phi^j(u_i, t) - \psi(u_{i-1})] dt \\ &= \frac{1}{\nu_i} \int_{t \in T_\gamma^j(u_i)} \rho_i^j(t)[\phi^j(u_i, t) - \psi(u_{i-1})] dt \end{aligned}$$

$$\begin{aligned}
(3.21) \quad & + \frac{(\alpha_{k_i} + \psi(u_{i-1}) - \psi(u_i))}{\nu_i} \left(\int_{t \in T_{\gamma}^j(u_i)^c} \frac{\phi^j(u_i, t) - \psi(u_{i-1})}{(\alpha_{k_i} + \psi(u_{i-1}) - \phi^j(u_i, t))^2} dt \right) \\
& \cong \psi(u_i) - \gamma - \psi(u_{i-1}) - \frac{(\alpha_{k_i} + \psi(u_{i-1}) - \psi(u_i))}{\nu_i} \\
& \quad \cdot \int_{t \in T_{\gamma}^j(u_i)^c} \frac{1}{\psi(u_{i-1}) - \phi^j(u_i, t)} dt \\
& \cong -\gamma + \psi(u_i) - \psi(u_{i-1}) - \frac{(\alpha_{k_i} + \psi(u_{i-1}) - \psi(u_i))}{\delta} \frac{1}{\gamma}.
\end{aligned}$$

Since $\psi(u_i) - \psi(u_{i-1}) \rightarrow 0$ and $\alpha_{k_i} \rightarrow 0$ as $i \rightarrow \infty$ and (3.21) is satisfied for any $\gamma > 0$, we obtain that

$$(3.22) \quad \frac{1}{\nu_i} \sum_{j=1}^q \int_0^1 \rho_i^j(t) [\phi^j(u_i, t) - \psi(u_{i-1})] dt \rightarrow 0, \quad \text{as } i \rightarrow \infty.$$

Hence, to complete our proof, we need only show that $\psi(u_i) \rightarrow 0$, as $i \rightarrow \infty$.

For contradiction, suppose that $\psi(\hat{u}) > 0$. We now have two possibilities. The first is that $2\psi(u_i) + f^0(u_{i-1}) - f^0(u_i) \rightarrow 0$ as $i \rightarrow \infty$, which implies that for all sufficiently large i , $f^0(u_i) \cong f^0(u_{i-1}) + \psi(\hat{u})$ and hence that $f^0(u_i) \rightarrow \infty$ as $i \rightarrow \infty$. Since the set G is bounded, this is clearly impossible. Hence, we consider the second possibility: There exists an infinite subset $K' \subset \mathbb{N}$ and a $\delta > 0$ such that $2\psi(u_i) + f^0(u_{i-1}) - f^0(u_i) \cong \delta$ for all $i \in K'$. In turn, this implies that $\int_0^1 \rho_i^0(t) dt \xrightarrow{K'} 0$ as $i \rightarrow \infty$. Furthermore, we may assume that $u_i \xrightarrow{K'} \sigma^* \in \bar{G}$, in the sense of control measures, as $i \rightarrow \infty$. Clearly, we must have $\bar{\psi}(\sigma^*) = \psi(\hat{u}) > 0$. In view of the above, (3.20), (3.21), and the continuity of $\bar{\theta}(\cdot)$, this implies that $\theta(\sigma^*) = 0$. However, this contradicts Assumption 3.1, and hence our proof is complete. \square

4. A special case. It is not at all uncommon for the set U in (2.1b) to be described in terms of convex inequalities, as follows:

$$(4.1) \quad U \triangleq \{z \in \mathbb{R}^m \mid s^l(z) \leq 0, l = 1, 2, \dots, q_3\},$$

where the $s^l: \mathbb{R}^m \rightarrow \mathbb{R}$ are all Lipschitz continuously differentiable convex functions.

If, for $j = q_1 + q_2 + 1, \dots, q_1 + q_2 + q_3$, we define the functions $\phi^j: L_{\infty,2} \times [0, 1] \rightarrow \mathbb{R}$ by

$$(4.2) \quad \phi^j(u, t) \triangleq s^{j-q_1-q_2}(u(t)),$$

we find that $\phi^j(\cdot, t)$ is differentiable on $L_{\infty,2}$, with $d_u \phi^j(u, t; u' - u) = \langle \nabla s^{j-q_1-q_2}(u(t)), u'(t) - u(t) \rangle$.

Now, let $q_4 \triangleq q_1 + q_2$ and $q_5 \triangleq q_1 + q_2 + q_3$; then (2.5a) becomes

$$(4.3a) \quad \mathbf{P}': \min \left\{ f^0(u) \mid \max_{t \in [0,1]} \phi^j(u, t) \leq 0, j = 1, 2, \dots, q_5, u \in L_{\infty,2} \right\},$$

or, in the even more compact form (see (2.5b)),

$$(4.3b) \quad \mathbf{P}': \min \{ f^0(u) \mid \psi^j(u) \leq 0, j = 1, 2, \dots, q_5, u \in L_{\infty,2} \},$$

where $\psi^j(u) \triangleq \max_{t \in [0,1]} \phi^j(u, t)$.

For problem \mathbf{P}' , letting $\psi(u) = \max \{ \psi^j(u), j = 1, \dots, q_5 \}$ and $\psi(u)_+ = \max \{ 0, \psi(u) \}$, we define the optimality function $\theta': L_{\infty,2} \rightarrow \mathbb{R}$ by

$$\begin{aligned}
(4.4) \quad & \theta'(u) \triangleq \min_{u' \in L_{\infty,2}} \left\{ \frac{1}{2} \|u' - u\|_2^2 + \max \{ df(u; u' - u) - 2\psi_+(u), \right. \\
& \quad \left. \phi^j(u, t) - \psi(u)_+ + d_u \phi^j(u, t; u' - u), \right. \\
& \quad \left. t \in [0, 1], j \in q_5 \} \right\}
\end{aligned}$$

which is a first-order convex approximation to the unifying function $F'(u'|u)$, defined by (2.6) with q and the functions $\phi^j(\cdot, \cdot)$ redefined as above.

We begin by establishing a relationship between the functions $\theta(\cdot)$ and $\theta'(\cdot)$.

THEOREM 4.1. *Suppose that*

$$(4.5) \quad \dot{U} \triangleq \{z \in U \mid s^j(z) < 0, j = 1, 2, \dots, q_3\},$$

where \dot{U} denotes the interior of U , and where \dot{U} is not empty. Then, for any $u \in G$, $\theta(u) = 0$ if and only if $\theta'(u) = 0$.

Proof. \Rightarrow Suppose that $\theta(\hat{u}) = 0$ but $\theta'(\hat{u}) < 0$. Since $\theta'(\hat{u}) < 0$, there exists a $\bar{u} \in L_{\infty,2}$ such that $F'(\bar{u}|\hat{u}) < 0$. If $\psi(\hat{u})_+ = 0$, then because of $F'(\bar{u}|\hat{u}) < 0$ we have that

$$(4.6a) \quad f^0(\bar{u}) - f^0(\hat{u}) - 2\psi(\hat{u})_+ = f^0(\bar{u}) - f^0(\hat{u}) < 0,$$

$$(4.6b) \quad \begin{aligned} \phi^j(\bar{u}, t) - \psi(\hat{u})_+ &= \phi^j(\bar{u}, t) < 0, \\ \forall t \in [0, 1], \text{ for } j &= 1, \dots, q_1 + q_2, \end{aligned}$$

$$(4.6c) \quad \begin{aligned} \phi^j(\bar{u}, t) - \psi(\hat{u})_+ &= \phi^j(\bar{u}, t) < 0, \\ \forall t \in [0, 1], \text{ for } j &= q_1 + q_2 + 1, \dots, q_1 + q_2 + q_3. \end{aligned}$$

By (4.6c) $\bar{u} \in G$. It now follows from (4.6a) and (4.6b) that $\theta(\hat{u}) < 0$, which contradicts our hypothesis.

Next, we need to prove that $\psi(\hat{u})_+ = 0$. Suppose that $\psi(\hat{u})_+ > 0$. Since $\hat{u} \in G$, $\phi^j(\hat{u}, t) \leq 0$ for all $j = q_5 \setminus q_4 \triangleq \{q_1 + q_2 + 1, \dots, q_1 + q_2 + q_3\}$ and for all $t \in [0, 1]$. Hence, $\psi(\hat{u})_+ > 0$ implies that there exist $j_0 \in q_4$ and $t_0 \in [0, 1]$ such that $\psi(\hat{u}) = \phi^{j_0}(\hat{u}, t_0) > 0$. However, in this case Assumption 3.1 ensures that $\theta(\hat{u}) < 0$, which is a contradiction, and therefore, $\psi(\hat{u})_+ = 0$.

Suppose that $\theta'(\hat{u}) = 0$. For the sake of contradiction, suppose that $\theta(\hat{u}) < 0$. Let $\xi(u|\hat{u})$ be defined by

$$(4.7a) \quad \begin{aligned} \xi(u|\hat{u}) &= \frac{1}{2}\|u - \hat{u}\|_2^2 + \max \{-2\psi(\hat{u})_+ + df^0(\hat{u}; u - \hat{u}), \\ &\quad \phi^j(\hat{u}, t) - \psi(\hat{u})_+ + d_u \phi^j(\hat{u}, t; u - \hat{u}), \\ &\quad t \in [0, 1], j = 1, \dots, q_1 + q_2\}. \end{aligned}$$

Since $\theta(\hat{u}) < 0$, there exists a $\bar{u} \in G$ such that

$$(4.7b) \quad \theta(\hat{u}) = \min_{u \in G} \xi(u|\hat{u}) = \xi(\bar{u}|\hat{u}) \triangleq -2\delta < 0.$$

For $\alpha \in (0, 1)$, let $u_\alpha \triangleq \hat{u} + \alpha(\bar{u} - \hat{u})$. Then $u_\alpha \in G$ for all $\alpha \in (0, 1)$ because the set G is convex. Since $-2\psi(\hat{u})_+ \leq 0$ and $\phi^j(\hat{u}, t) - \psi(\hat{u})_+ \leq 0$ for all j , we have that, for all $\alpha \in (0, 1)$,

$$(4.7c) \quad \xi(\bar{u}|\hat{u}) \leq \xi(u_\alpha|\hat{u}) \leq -2\alpha\delta < 0.$$

Let \tilde{u} be any control such that $\tilde{u}(t) \in \dot{U}$ for all $t \in [0, 1]$. Then, by (4.5), there exist $\bar{\delta}_j$ such that $\max_{t \in [0, 1]} \phi^j(\tilde{u}, t) = -\bar{\delta}_j < 0$ for all $j \in q_5 \setminus q_4$. Let $u'_{\alpha, \beta} \triangleq (1 - \alpha)\hat{u} + \alpha(\bar{u} + \beta(\tilde{u} - \bar{u})) = u_\alpha + \alpha\beta(\tilde{u} - \bar{u})$, where $\beta \in (0, 1)$. Obviously, $u'_{\alpha, \beta} \in G$. By the continuity of $\xi(\cdot|\hat{u})$, for any $\alpha \in (0, 1)$, there exists a $\beta_\alpha \in (0, 1)$ such that

$$(4.7d) \quad \theta(\hat{u}) \leq \xi(u'_{\alpha, \beta}|\hat{u}) \leq -\alpha\delta < 0, \quad \forall 0 < \beta \leq \beta_\alpha.$$

Since the functions $\phi^j(\cdot, t)$ are convex for all $j \in q_5 \setminus q_4$ and for all $t \in [0, 1]$, we conclude that, for all $j \in q_5 \setminus q_4$, for all $t \in [0, 1]$, for all $\alpha \in (0, 1)$, and for all $\beta \in (0, 1)$,

$$(4.8a) \quad \begin{aligned} \phi^j(u'_{\alpha, \beta}, t) &\leq (1 - \alpha)\phi^j(\hat{u}, t) + \alpha\phi^j(\bar{u} + \beta(\tilde{u} - \bar{u}), t) \\ &\leq (1 - \alpha)\phi^j(\hat{u}, t) + \alpha\{(1 - \beta)\phi^j(\bar{u}, t) + \beta\phi^j(\tilde{u}, t)\} \\ &\leq -\alpha\beta\bar{\delta}_j < 0. \end{aligned}$$

The last inequality is valid because $\phi^j(\hat{u}, t) \leq 0$, $\phi^j(\bar{u}, t) \leq 0$, and $\phi^j(\tilde{u}, t) \leq -\bar{\delta}_j < 0$. Also, because all the gradients $\nabla_u \phi^j(\cdot, t)$ are Lipschitz continuous,

$$\begin{aligned}
 \phi^j(u'_{\alpha,\beta}, t) &= \phi^j(\hat{u}, t) + \langle \nabla_u \phi^j(\hat{u}, t), u'_{\alpha,\beta}(t) - \hat{u}(t) \rangle \\
 &\quad + \int_0^1 \langle (\nabla_u \phi^j(\hat{u} + s(u'_{\alpha,\beta} - \hat{u}), t) - \nabla_u \phi^j(\hat{u}, t)), u'_{\alpha,\beta} - \hat{u} \rangle ds \\
 &\geq \phi^j(\hat{u}, t) + \langle \nabla_u \phi^j(\hat{u}, t), u'_{\alpha,\beta}(t) - \hat{u}(t) \rangle \\
 &\quad - L_j \alpha^2 / 2 (\|\bar{u}(t) - \hat{u}(t)\|^2 + \beta^2 \|\tilde{u}(t) - \bar{u}(t)\|^2),
 \end{aligned}
 \tag{4.8b}$$

where $L_j \triangleq \max_{t \in [0,1]} L_j(t)$ and $L_j(t)$ is a Lipschitz constant of $\nabla_u \phi^j(\cdot, t)$. Combining (4.8a) with (4.8b), we can conclude that there exist $\alpha_0, \beta_0 \in (0, 1)$ such that, for all $0 < \alpha \leq \alpha_0$, for all $0 < \beta \leq \beta_0$, for all $t \in [0, 1]$, and for all $j \in q_5 \setminus q_4$,

$$\begin{aligned}
 &\frac{1}{2} \|u'_{\alpha,\beta}(t) - \hat{u}(t)\|^2 + \phi^j(\hat{u}, t) + \langle \nabla \phi^j(\hat{u}, t), u'_{\alpha,\beta}(t) - \hat{u}(t) \rangle \\
 &\leq -\alpha \beta \bar{\delta}_j / 2 < 0.
 \end{aligned}
 \tag{4.8c}$$

Since (4.7d) and (4.8c) imply that, given $0 < \alpha \leq \alpha_0$, for all $0 < \beta \leq \min\{\beta_\alpha, \beta_0\}$,

$$\theta'(\hat{u}) \leq \max\{-\alpha \delta, -\alpha \beta \bar{\delta}_j / 2, j \in q_5 \setminus q_4\} < 0,
 \tag{4.8d}$$

we obtain a contradiction of our hypothesis, and hence our proof is complete. \square

Clearly, Algorithm 3.1 is applicable to \mathbf{P}' , and it may be initialized with a control that is not in G . However, we must amend Assumption 3.1 as follows. Since relaxed controls must be associated with bounded controls, we introduce an arbitrarily large compact set $U^* \subset \mathbb{R}^n$, and we define G^* by (2.1b) with U replaced by U^* . We denote the corresponding set of relaxed controls by \bar{G}^* .

Assumption 4.1. (i) The \bar{G}^* -closure of the set $\{u \in G^* \mid \psi(u) \leq 0\}$ is equal to the \bar{G}^* -closure of its interior. (ii) For all $u \in G^*$ such that $\psi(u) > 0$, $\theta'(u) < 0$.

At this point, the following result should be obvious.

THEOREM 4.2. (i) Suppose that Algorithm 3.1 is applied to problem \mathbf{P}' and jams up in the loop between Steps 1 and 2 at the control u_{i_0} . Then $\psi(u_{i_0}) \leq 0$ and $\theta'(u_{i_0}) = \theta(u_{i_0}) = 0$.

(ii) Suppose that $\{u_i\}_{i=0}^\infty$ is a sequence of controls constructed by Algorithm 3.1 in solving \mathbf{P}' . If this sequence has an accumulation point $\hat{u} \in L_{\infty,2}$, then $\psi(\hat{u}) \leq 0$ (so that $\hat{u} \in G$) and $\theta'(\hat{u}) = \theta(\hat{u}) = 0$.

There is considerable programming convenience in using the formulation \mathbf{P}' over \mathbf{P} when possible. Our computational results in the next section show that the use of the formulation \mathbf{P}' does not result in any penalty in terms of computing times.

5. Numerical results. We now present two examples that illustrate the performance of Algorithm 3.1. In our experiments, the computations in Step 1 of Algorithm 3.1 were carried out by using Algorithm A in [21], which is of the Gauss-Newton type. The sequence $\{\alpha_k\}_{k=0}^\infty$ was defined by $\alpha_{k+1} = \alpha_k / 1.1$ with $\alpha_0 = 0.005$. Our experiments suggest that larger values of α_0 result in an increase in the number of iterations needed to solve a problem.

Example 5.1. Our first problem is a minimum-time brachistochrone problem with a state-variable inequality constraint, described in [4]. We treat this problem in fixed-time scaled form, where the scale variable T corresponds to the actual final time.

$$\mathbf{P}: \min_{\gamma \in L_{\infty,2}} \left\{ \frac{1}{2} T^2 \mid \max_{t \in [0,1]} \phi^j(\gamma, t) \leq 0, j = 1, 2, \forall t \in [0, 1] \right\},
 \tag{5.1a}$$

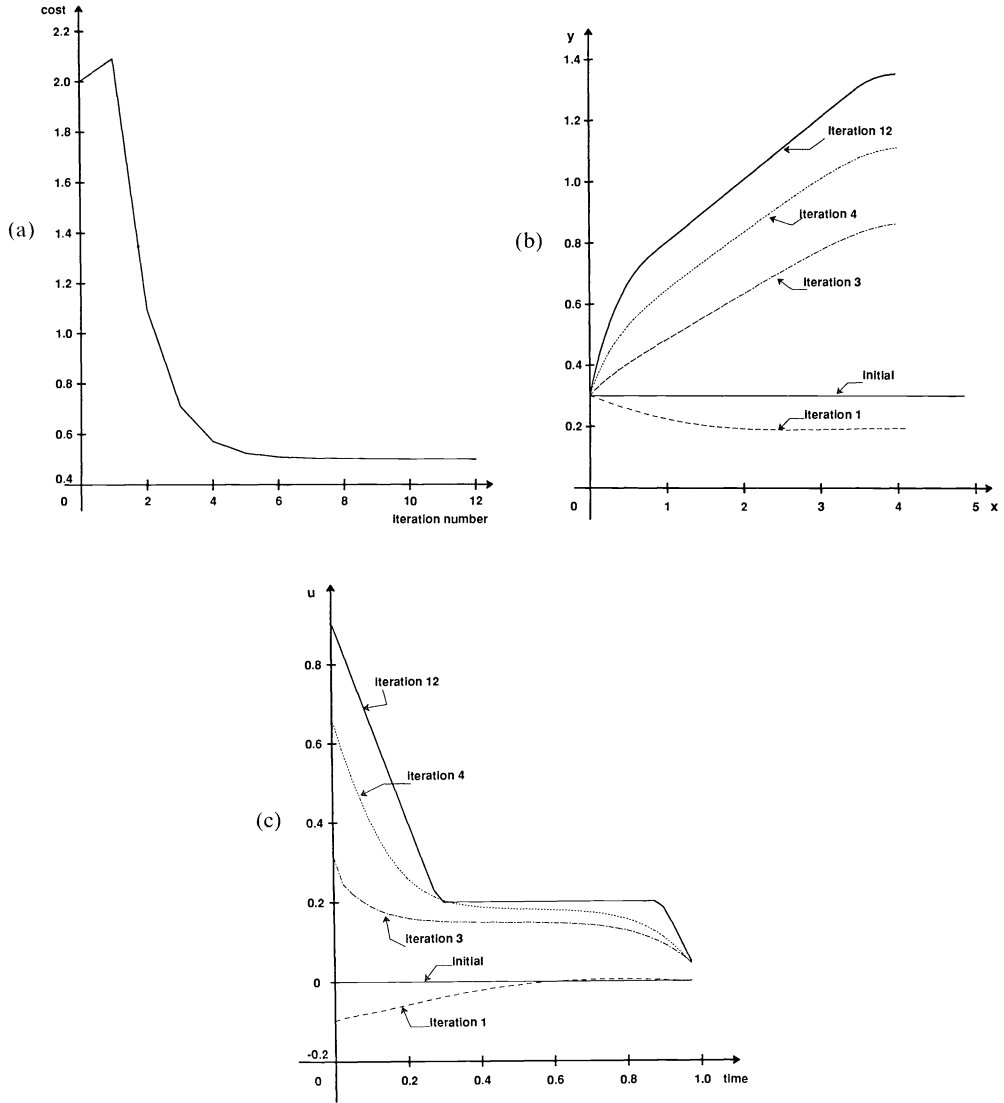


FIG 1. (a) Cost versus iteration number, (b) state-space trajectories, and (c) controls at various iterations for Example 5.1 with set (i) of initial conditions.

where

$$(5.1b) \quad \phi^1(\gamma, t) = y(t) - x(t) \tan \theta - h,$$

$$(5.1c) \quad \phi^2(\gamma, t) = \frac{1}{2}(x(1) - l)^2 - \xi,$$

where horizontal distance x and vertical distance y are defined by

$$(5.2a) \quad \dot{x}(t) = T\sqrt{2gy(t)} \cos \gamma(t),$$

$$(5.2b) \quad \dot{y}(t) = T\sqrt{2gy(t)} \sin \gamma(t),$$

where g is the acceleration due to gravity; γ is the path angle to the horizontal; θ and h are constants; and ξ is a small tolerance by which we are willing to relax the requirement $x(1) = l$.

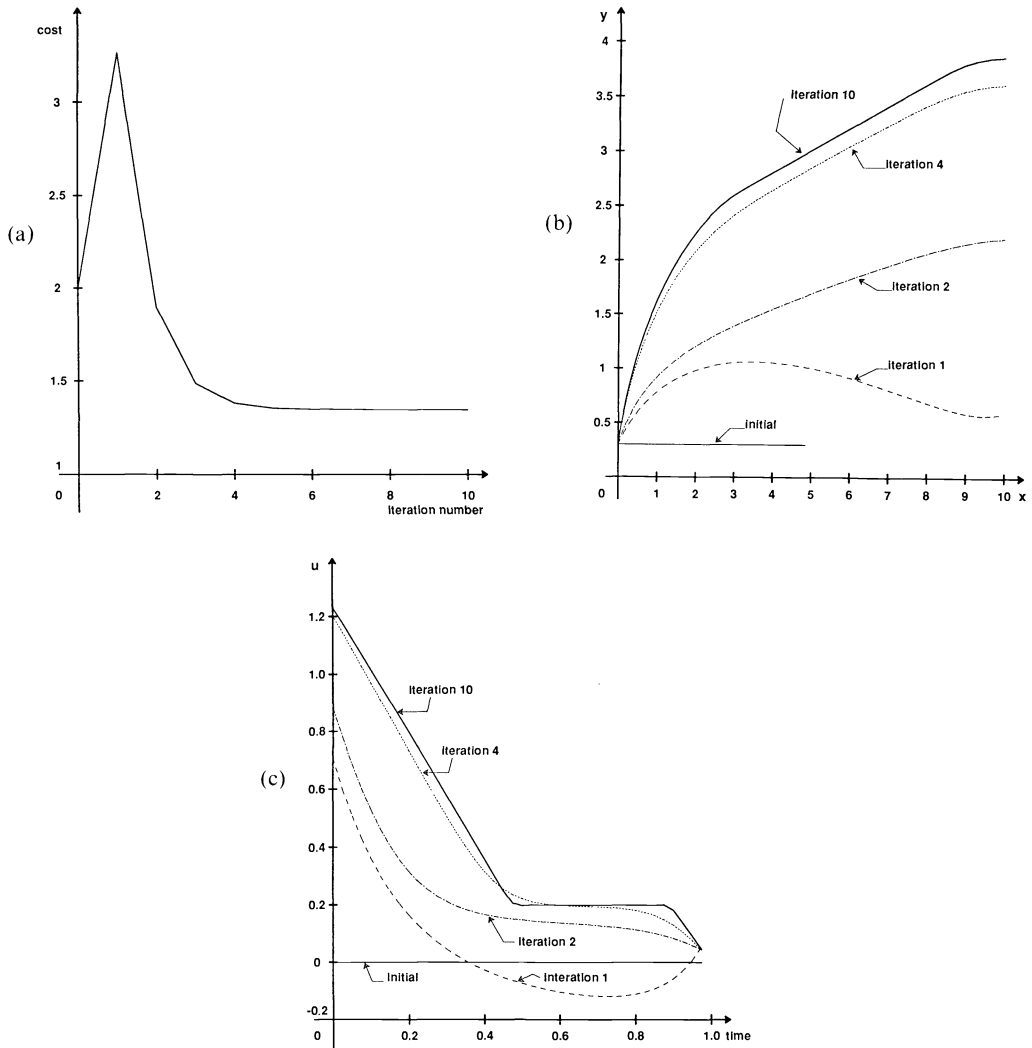


FIG. 2. (a) Cost versus iteration number, (b) state-space trajectories, and (c) controls at various iterations for Example 5.1 with set (ii) of initial conditions.

Since the system equations (5.2a) and (5.2b) cannot be integrated explicitly, we must use a numerical integration scheme, which discretizes the time interval $[0, 1]$. The discretization may be either fixed or variable. We used 40 uniformly spaced points in conjunction with the Runge-Kutta second-order method [25]. The gradients used were those corresponding to the discretized dynamics imposed by the integration scheme. The results that we obtained converge to the expected results, obtained analytically in [4, p. 120]. We stopped our computations when the constraints were satisfied and the difference in the cost values between successive iterations was less than 1×10^{-5} .

We used the following two sets of initial conditions: (i) $(x(0), y(0), T) = (0.0, 0.3, 2.0)$, $\theta = 0.2$, $h = 0.6$, $l = 4.0$, and $\xi = 0.0005$; and (ii) $(x(0), y(0), T) = (0.0, 1.0, 2.0)$, $\theta = 0.2$, $h = 2.0$, $l = 10.0$, and $\xi = 0.0005$.

Figure 1(a) presents a plot of the values $f^0(u_i)$ versus iteration number i . Figures 1(b) and 1(c) show state-space trajectories and inputs at various iterations, respectively,

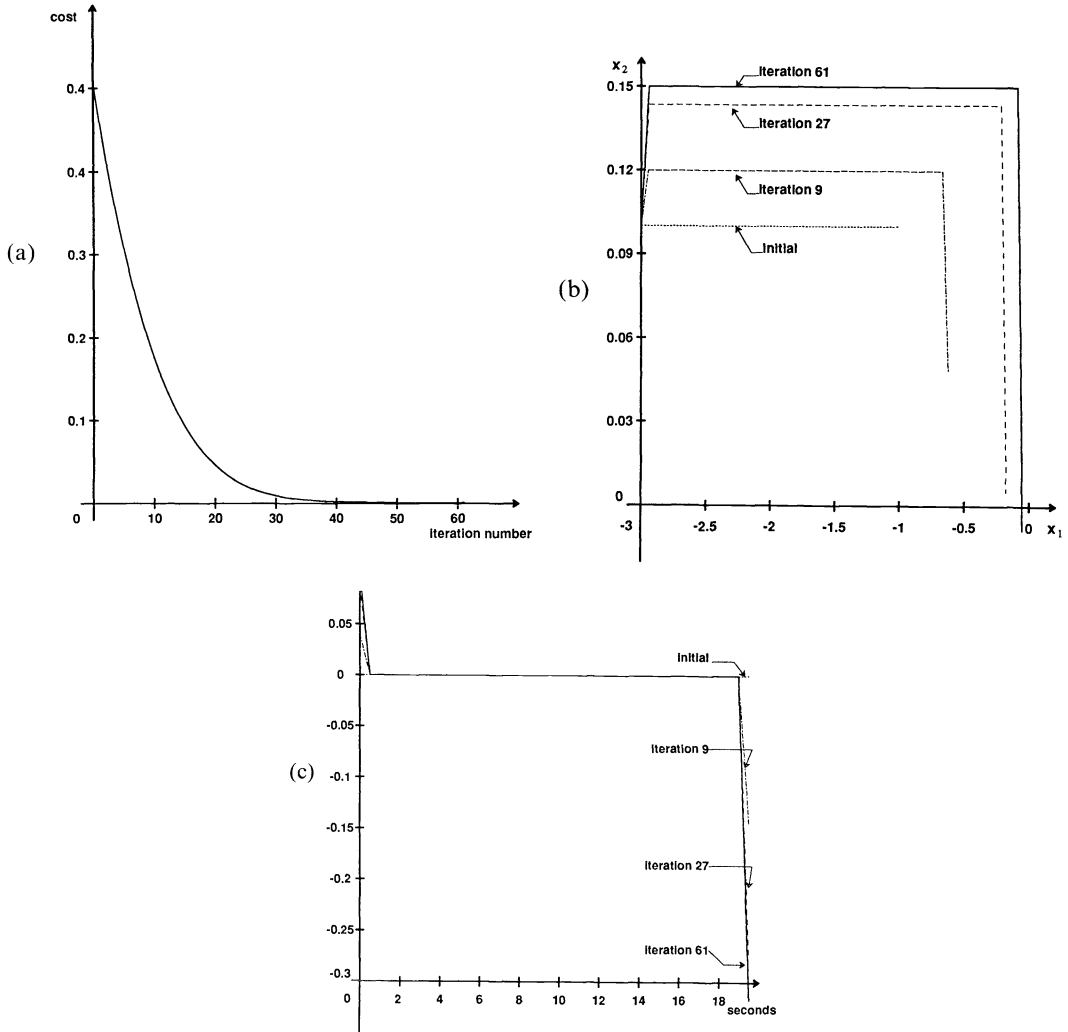


FIG. 3. (a) Cost versus iteration number, (b) state-space trajectories, and (c) controls at various iterations for Example 5.2 using penalized control.

obtained by using the first set of initial conditions. The value of $\frac{1}{2}T^2$ is determined to be 0.99971 after twelve iterations. The results of the computations using the second set of initial conditions are shown in Figs. 2(a)-2(c). The final value of $\frac{1}{2}T^2$, 1.63998 seconds, is obtained after ten iterations.

Example 5.2. Our second problem is a fixed-time minimum final error problem with a state-variable inequality constraint and bounds on the control

$$(5.3a) \quad \mathbf{P}: \min_{u \in G} \left\{ \frac{1}{2} \|x(1)\|^2 \mid x^2(t) - l \leq 0, \forall t \in [0, 1] \right\},$$

where the state is determined by the scaled differential equation

$$(5.3b) \quad \dot{x}(t) = \begin{pmatrix} \dot{x}^1(t) \\ \dot{x}^2(t) \end{pmatrix} = T \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + T \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t),$$

with $T > 0$ the actual final time. The input $u(\cdot)$ is scalar-valued and $u \in G \triangleq \{u \in L_{\infty,2} \mid \|u\|_{\infty} \leq 1.0\}$. We use the initial state given by $x(0) = (-3.0, 0.1)^T$. We set $l = 0.15$

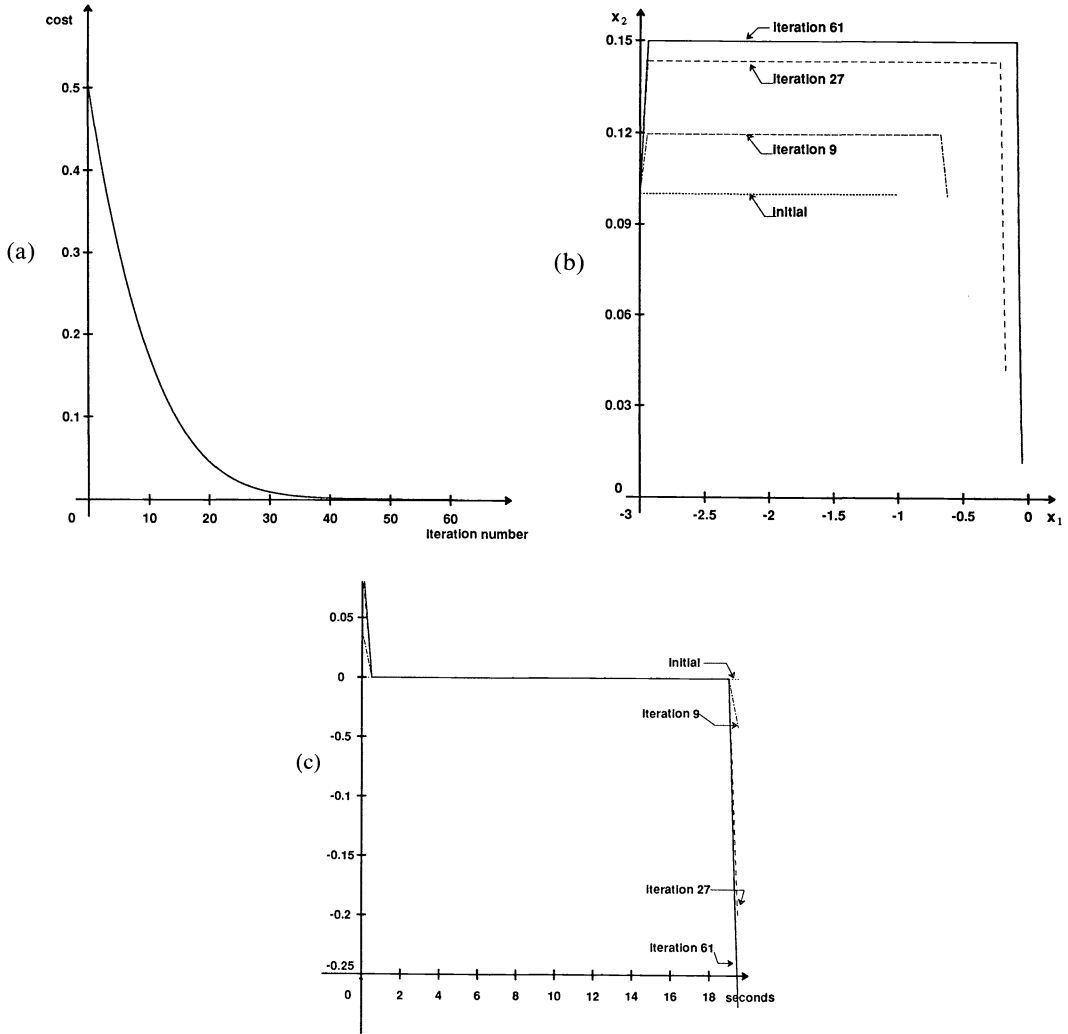


FIG. 4. (a) Cost versus iteration number, (b) state-space trajectories, and (c) controls at various iterations for Example 5.2 using nonpenalized control.

and $T = 20$. With $s(\cdot)$ defined by $s(t) = u(t)^2 - 1.0$, problem **P** (5.3a) becomes

$$(5.4a) \quad \mathbf{P}': \min_{u \in L_{\infty,2}} \left\{ \frac{1}{2} \|x(1)\|^2 \right\} \max_{t \in [0,1]} \phi^j(u, t) \leq 0, \quad j = 1, 2,$$

where

$$(5.4b) \quad \phi^1(u, t) = x^2(t) - l,$$

$$(5.4c) \quad \phi^2(u, t) = s(t).$$

We applied Algorithm 3.1 to both problems **P** and **P'**.

Figures 3(a)–3(c) present plots of the cost $f^0(u_i)$ versus iteration number i , as well as corresponding state-space trajectories and inputs at various iterations obtained by using formulation (2.1a). Similar results from using formulation (4.3b) are shown in Figs. 4(a)–4(c). As we can see, the results obtained are almost identical.

It is clear from our experimental results that Algorithm 3.1 is effective in solving optimal control problems with continuum state and control constraints. Also, when we have a special description of the set of controls, we can take advantage of it without any penalty.

6. Conclusion. We have presented two versions of a phase I-phase II method of centers algorithm for the solution of optimal-control problems with control and state-space constraints. The computational advantages of these algorithms derive from the fact that we used barrier functions for defining an approximate center to be computed at each iteration. Although, at first glance, the algorithms appear to have potential for failure due to ill conditioning, preliminary computational results show that this is not so and that, in fact, the algorithms are highly effective. This observation agrees with the numerical results reported in [21] for a related algorithm that solves semi-infinite minimax problems.

REFERENCES

- [1] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math. 16 (1966), pp. 1-3.
- [2] T. E. BAKER AND E. POLAK, *On the Optimal Control of Systems Described by Evolution Equations*, Memo UCB/ERL M89/113, Electronics Research Lab., Univ. of California, Berkeley, CA, September 1989.
- [3] D. P. BERTSEKAS, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 174-184.
- [4] A. E. BRYSON JR. AND Y.-C. HO, *Applied Optimal Control; Optimization, Estimation, and Control*, revised ed., Hemisphere, New York, 1975.
- [5] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist., 3 (1956), pp. 95-110.
- [6] W. W. HAGER, *Multiplier methods for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061-1080.
- [7] P. HUARD, *Programmation mathématique convexe*, Rev. Fr. Inf. Rech. Oper., 7 (1968), pp. 43-59.
- [8] R. KLESSIG AND E. POLAK, *An adaptive algorithm for unconstrained optimization with applications to optimal control*, SIAM J. Control, 11 (1973), pp. 80-94.
- [9] K. C. P. MACHIELSON, *Numerical Solution of Optimal Control Problems with State Constraints by Sequential Quadratic Programming in Function Space*, Centrum voor Wiskunde en Informatica, Amsterdam, 1988.
- [10] H. MAURER AND W. GILLESSEN, *Application of multiple shooting to the numerical solution of optimal control problems with bounded state variables*, Computing, 15 (1975), pp. 105-126.
- [11] D. Q. MAYNE AND E. POLAK, *A feasible directions algorithm for optimal control problems with terminal inequality constraints*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 741-751.
- [12] ———, *An exact penalty function algorithm for control problems with state and control constraints*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 380-387.
- [13] ———, *An exact penalty function algorithm for optimal control problems with control and terminal equality constraints, part 1*, J. Optim. Theory Appl., 32 (1980), pp. 211-246.
- [14] ———, *An exact penalty function algorithm for optimal control problems with control and terminal equality constraints, part 2*, J. Optim. Theory Appl., 32 (1980), pp. 345-363.
- [15] A. MIELE AND A. K. WU, *Sequential conjugate gradient restoration algorithm for optimal control problems with nondifferentiable constraints and general boundary conditions, part I, theory*, Opt. Control Appl. Methods, 1 (1980), pp. 69-88.
- [16] R. MIFFLIN, *Rates of convergence for a method of centers algorithm*, J. Optim. Theory Appl., 18 (1976), pp. 199-228.
- [17] O. PIRONNEAU AND E. POLAK, *On the rate of convergence of certain methods of centers*, Math. Programming, 2 (1972), pp. 230-258.
- [18] ———, *A dual method for optimal control problems with initial and final boundary constraints*, SIAM J. Control, 11 (1973), pp. 534-549.

- [19] E. POLAK, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–91.
- [20] E. POLAK AND L. HE, *A unified phase I-phase II method of feasible directions for semi-infinite optimization*, Memo UCB/ERL M89/7, Electronics Research Lab., Univ. of California, Berkeley, CA, February 1989; J. Optim. Theory Appl., 69 (1991), pp. 83–107.
- [21] E. POLAK, J. HIGGINS, AND D. Q. MAYNE, *A barrier function method for minimax problems*, Memo UCB/ERL M88/64, Electronics Research Lab., Univ. of California, Berkeley, CA, October 1988; Math. Programming, 54 (1992), pp. 155–176.
- [22] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1972.
- [23] E. POLAK AND D. Q. MAYNE, *First order, strong variations algorithms for optimal control problems with terminal inequality constraints*, J. Optim. Theory Appl., 16 (1975), pp. 303–325.
- [24] ———, *An algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 184–193.
- [25] A. RALSTON, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965.
- [26] J. WARGA, *Optimal Control of Differential Equations and Functional Equations*, Academic Press, New York, 1972.
- [27] ———, *Steepest descent with relaxed controls*, SIAM J. Control Optim., 15 (1977), pp. 674–682.
- [28] ———, *Iterative procedures for constrained and unilateral optimization problems*, SIAM J. Control Optim., 20 (1982), pp. 360–367.
- [29] ———, *Iterative optimization with equality constraints*, Math. Oper. Res., 9 (1984), pp. 592–605.
- [30] L. J. WILLIAMSON AND E. POLAK, *Relaxed controls and the convergence of optimal control algorithms*, SIAM J. Control Optim., 14 (1976), pp. 737–757.

NEARLY OPTIMAL CONTROLS FOR STOCHASTIC ERGODIC PROBLEMS WITH PARTIAL OBSERVATION*

WOLFGANG J. RUNGGLADIER† AND ŁUKASZ STETTNER‡

Abstract. This paper considers the problem of constructing nearly optimal controls for discrete-time, infinite horizon, stochastic control problems under partial observations, with the average cost criterion.

Key words. stochastic control, partial observations, ergodic control, invariant measures, nearly optimal controls, approximation methods

AMS(MOS) subject classifications. primary 93E20; secondary 93E11, 93E25

1. Introduction. Stochastic control problems with partial observation have considerable importance in applications, and a large body of literature is available concerning this subject (see, e.g., [5] and references therein). Nevertheless, while many results are available concerning existence of optimal controls, little is known concerning the construction of an optimal, or at least nearly optimal, control. Here we address such a problem in the case of infinite horizon with the average cost criterion and partial observations of the state. The construction of nearly optimal controls is closely related to approximations of stochastic control problems, and in this area some pioneering work is due to Kushner (see [8]; see also the more recent survey [9]). In line with previous work [2] on using approximation methods to derive nearly optimal controls, in a first paper [12] we studied the case of infinite horizon with discounting and partial observations of the state. Here we again study the infinite horizon problem, but with the average cost criterion and partial observations of the state, which required a rather different approach and for which, to our knowledge, nothing has yet appeared in the literature (see [5], where, in the case of partially observed diffusions, open problem 6 is such a problem).

We consider a rather general setting of the problem, where the state process is Markov, characterized by its transition kernel, and evolves on a locally compact state space. On the other hand, we treat only the discrete-time case, but this case arises in many circumstances and is natural when the observations, and therefore also the controls, are taken at discrete-time points.

As admissible controls, we take those corresponding to the so-called separated problem, namely, those that depend on past and present observations only through the filter values. We do not address the problem of whether the optimal control for the separated problem is optimal also when the controls are simply adapted to the σ -field generated by the observations. Such a fact is true in many situations (see, e.g., [3]; see also [12] for a proof in the continuous-time, infinite horizon case with discounting, when, besides continuously acting controls, also impulsive control and stopping are allowed), and here we simply accept it, concentrating our efforts on the construction of a method to obtain a control that is nearly optimal with respect to

* Received by the editors October 29, 1990; accepted for publication (in revised form) September 2, 1991.

† Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy, and LADSEB-CNR, Padova, Italy.

‡ Institute of Mathematics, Polish Academy of Sciences, ul. Śniadeckich 8, 00-950 Warsaw, Poland. The work of this author was performed during a stay in Padova, sponsored by GNAFA/CNR, the Italian National Research Council.

such a class of admissible controls. The class of controls that we consider, and the assumptions that we impose, are mainly motivated by the need to have a unique invariant measure for the controlled filtering process.

The method consists of two basic parts: The first part—essentially § 3—consists of a procedure to determine a nearly optimal control function that, when applied to the true filter values, yields nearly optimal controls. Since generally the true filter values cannot be computed in practice, in the second part—essentially § 4—we introduce a computable approximate filter and show that, when applying the nearly optimal control functions of the first part to the approximate filter values, we still obtain nearly optimal controls.

In § 1.1 we give a precise formulation of the problem. In § 1.2 we state and explain our assumptions. Section 1.3 is devoted to some preliminary results, including results on measure transformations that allow the problem to be cast in its proper frame. Section 2 contains some fundamental results for the rest of the paper, namely, on convergence of invariant measures that underlie many more specific results, as well as on approximations of controls. Section 3 describes the first part of our approximation method, namely, the construction of a control function that, when applied to the true filter values, yields nearly optimal controls for the original problem. Section 3.3 summarizes the algorithmic aspects of our (approximation) procedure. Finally, § 4 concerns the definition of a computable approximate filter process and the proof that, when evaluating the nearly optimal control function—constructed in § 3—at the approximating filter values, we still obtain nearly optimal controls for the original problem.

Concerning notation, we sometimes use $\mathcal{B}(A)$ to denote both the Borel sets in A as well as the (bounded) Borel functions on A ; the context should create no ambiguity.

1.1. Problem formulation. On a given probability space (Ω, \mathcal{F}, P) , consider a controlled Markov process (x_i) , $i = 1, 2, \dots$, with values in a locally compact separable state space E . Assume that (x_i) has initial law μ and denote by $P^{u_i}(x, d\eta)$ its transition kernel in the generic period i , where u_i represents the control that takes values in a set of control parameters $U \subset \mathcal{R}^k$, which is compact and such that $\{0\} \in U$. The process (x_i) is only partially observed through observations (y_i) , $y_i \in \mathcal{R}^d$, defined by

$$(1.1) \quad y_i = h(x_i) + w_i, \quad i = 1, 2, \dots,$$

where $h \in C(E, \mathcal{R}^d)$, the space of continuous bounded functions from E into \mathcal{R}^d , and (w_i) are independently and identically distributed d -dimensional standard Gaussian random variables, independent of x_k for $k \leq i$. We assume that each u_i is adapted to the observation σ -algebra $Y^i := \sigma\{y_1, \dots, y_i\}$. Given a bounded Borel function ϕ on E , now define a process $\pi_i^{\mu, u}(\cdot)$ with values in the space $\mathcal{P}(E)$ of probability measures on E , endowed with the topology of weak convergence, as follows:

$$(1.2) \quad \pi_0^{\mu, u}(\phi) = \mu(\phi), \quad \pi_i^{\mu, u}(\phi) = E_\mu\{\phi(x_i) | Y^i\},$$

where E_μ stands for the expectation, given the initial law μ for the process (x_i) that is controlled by a law u to be defined below and where we use $\pi(\phi)$ to denote the integral of a function $\phi(x)$ with respect to the measure $\pi(dx)$. We call $(\pi_i^{\mu, u})$ the filtering process corresponding to the controlled process (x_i) with observations (1.1). This process can easily be seen to be Markov; its transition kernel will be denoted (see (1.33) below) by $\Pi^u(\mu, \Lambda)$, with $\Lambda \in \mathcal{B}(\mathcal{P}(E))$ the σ -field of Borel subsets of $\mathcal{P}(E)$.

We are now more specific as to the sense in which we consider u_i to be Y^i -adapted: We define the class \mathcal{A} of admissible control functions as the class

$$\mathcal{A} = \{u : \mathcal{P}(E) \rightarrow U; u \text{ is continuous}\}.$$

By analogy to known results in the discounted infinite horizon case concerning the so-called separated problem (for a proof, see, e.g., [12, Thm. 3]), in the first part of the paper, more precisely in §§ 1–3, we then consider Y^i -adapted controls u_i of the form

$$(1.3) \quad u_i = u(\pi_i^{\mu, u}), \quad u \in \mathcal{A}.$$

Since generally the true filtering process cannot be computed, in the last part of the paper, essentially § 4, we consider Y^i -adapted controls of the form (1.3), where $\pi_i^{\mu, u}$ is replaced by a computable approximate filtering process.

Being interested in ergodic stochastic control, given a continuous and bounded function $c(x, v) : E \times U \rightarrow \mathcal{R}$, we consider as an objective function to be minimized the following functional:

$$(1.4) \quad J_\mu(u) = \limsup_{n \rightarrow \infty} n^{-1} \sum_{i=0}^{n-1} E_\mu\{c(x_i, u(\pi_i^{\mu, u}))\}.$$

The minimization is over the class \mathcal{A} of admissible control functions. In §§ 2 and 3, we describe a feasible method for the construction of a nearly optimal control function $u(\cdot)$ for the cost functional (1.4) in the class \mathcal{A} as well as in restricted subclasses to be defined later.

Given a nearly optimal control function \hat{u} , we are still not able to apply the controls of the form (1.3) since the filtering process $(\pi_i^{\mu, u})$ cannot be computed explicitly. Therefore, in § 4 we define a computable approximating filtering process $(\pi_i^{m(\mu)})$ taking values in a finite-dimensional simplex S^m ; we also define a mapping $\tilde{\mathcal{J}}_m$ from \mathcal{A} into functions defined on S^m such that, given a nearly optimal control function $\hat{u} \in \mathcal{A}$, the controls

$$(1.5) \quad u_i = \tilde{\mathcal{J}}_m \hat{u}(\pi_i^{m(\mu)})$$

are nearly optimal, i.e.,

$$\limsup_{n \rightarrow \infty} n^{-1} \sum_{i=0}^{n-1} E_\mu\{c(x_i, u_i)\} \leq \inf_{u \in \mathcal{A}} J_\mu(u) + \varepsilon,$$

thus completing the description of a feasible method for the construction of nearly optimal controls for our ergodic control problem.

1.2. Basic assumptions and their discussion. We make the following assumptions:

- (A1) For fixed $v \in U$, the transition kernel $P^v(x, \cdot)$ is Feller.
- (A2) If $U \ni v_m \rightarrow v$, then, letting \Rightarrow denote weak convergence, $P^{v_m}(x, \cdot) \Rightarrow P^v(x, \cdot)$ uniformly for x from compact subsets of E , i.e., for any $f \in C(E)$, $P^{v_m}f(x) \rightarrow P^vf(x)$, uniformly on compact subsets of E .
- (A3) For each open set $\mathcal{O} \subset E$, $v \in U$, $x \in E$, $P^v(x, \mathcal{O}) > 0$.
- (A4) There exists $\eta(\cdot) \in \mathcal{P}(E)$ such that for all $x \in E$ we have $P^v(x, \cdot) = \eta(\cdot)$ for $v = 0$.

Remark 1.2.1. Assumptions (A1) and (A2) guarantee (see Proposition 1.2(ii), below) the Feller property of the filtering process $(\pi_i^{\mu, u})$ and seem to be natural. Assumption (A3) corresponds to the nondegeneracy of the controlled transition kernel of the state process. Assumption (A4) might appear restrictive, but is satisfied in a

variety of applied models (see, e.g., [1]); more generally, it holds in the following example.

Example (EX). Let $E = \mathcal{R}^n$ and let x_i satisfy

$$(1.6) \quad x_{i+1} = f(x_i, u_i) + g(x_i, u_i)v_i, \quad x_0 = x$$

where $f(x, u)$, $g(x, u)$ are continuous functions such that for all $x \in E$ we have $f(x, 0) = \text{const}$, $g(x, 0) = c \cdot I$ ($c \neq 0$), and where (v_i) is an independently and identically distributed sequence of standard Gaussian random vectors, independent of (w_i) in (1.1).

Note that, for the example, (A1)–(A3) are also satisfied.

We now make further assumptions that depend on whether E is locally compact without being compact. In the case where E is locally compact, but noncompact, we make the following further assumptions.

(B1) There exists $j \in \{1, 2, \dots, d\}$ such that the j th component $h^j(x)$ of $h(x)$ has a limit at “ ∞ ” and attains at “ ∞ ” either its strong maximum or strong minimum. More precisely, letting

$$(1.7) \quad K_n = \{x \in E \mid \rho(x, \bar{x}) \leq n\},$$

where ρ is a metric on E compatible with the topology and \bar{x} is a fixed element of E , we either have

$$(1.8) \quad \sup_{x \in K_n} h^j(x) < \sup_{x \in E} h^j(x) \quad \text{for } n = 1, 2, \dots$$

or

$$(1.9) \quad \inf_{x \in K_n} h^j(x) > \inf_{x \in E} h^j(x).$$

(B2) For each $\varepsilon > 0$, there exists $u_\varepsilon \in \mathcal{A}$ and a unique invariant measure Φ^{u_ε} corresponding to the transition operator $\Pi^{u_\varepsilon}(\mu, \cdot)$, such that u_ε is ε -optimal in the sense that, for a given initial law μ of the state process (x_i) , we have

$$(1.10) \quad \int_{\mathcal{P}(E)} \int_E c(x, u_\varepsilon(v)) \nu(dx) \Phi^{u_\varepsilon}(d\nu) = J_\mu(u_\varepsilon) \leq \inf_{u \in \mathcal{A}} J_\mu(u) + \varepsilon.$$

(B3) There exists an initial law μ such that the family of Cesaro averages

$$\left\{ \frac{1}{t} \sum_{i=0}^{t-1} (\Pi^u)^i(\mu, \cdot), \quad u \in \mathcal{A}, \quad t = 1, 2, \dots \right\}$$

is tight.

(B4) For any compact set $K \subset E$, there exists $\alpha > 0$ such that

$$\inf_{v \in U} \inf_{x \in E} P^v(x, K^c) \geq \alpha.$$

Remark 1.2.2. Although assumption (B2) contains an usual requirement in ergodic control (see, e.g., [10; § 6]), its verification in general cases seems to be an open problem (in particular, concerning the existence of a unique invariant measure). This problem is studied in [6], but for a very particular partially observed model. On the other hand, if we restrict the class of admissible control functions \mathcal{A} to a subclass $\bar{\mathcal{A}} \subset \mathcal{A}$, then, given the other assumptions, we may hope to be able to obtain uniqueness of the invariant measure for $(\pi_i^{\mu, u})$. This is indeed the case and, to define the subclass $\bar{\mathcal{A}}$, let $r: [0, 1] \rightarrow [0, 1]$ be a continuous nondecreasing function such that, with $0 < b < c < 1$,

$$(1.11) \quad r(x) = \begin{cases} 0 & \text{for } x < b, \\ 1 & \text{for } x > c. \end{cases}$$

Furthermore, let $\psi_n(x) \in C(E)$ with values in $[0, 1]$ be a given function satisfying

$$(1.12) \quad \psi_n(x) = \begin{cases} 1 & \text{for } x \in K_{n-1}, \\ 0 & \text{for } x \in E \setminus K_n. \end{cases}$$

Now define

$$(1.13) \quad \bar{\mathcal{A}} = \bigcup_{n=1}^{\infty} \bar{\mathcal{A}}_n \quad \text{with } \bar{\mathcal{A}}_n = \{u \in \mathcal{A} \mid u(v) = \tilde{u}(v)r(v(\psi_n)) \text{ for } \tilde{u} \in \mathcal{A}\};$$

i.e., $\bar{\mathcal{A}}$ consists of control functions in \mathcal{A} that are equal to zero in a weak neighborhood of “ ∞ .” It follows from Corollary 2.4, below, that, given the other assumptions, (B2) is automatically satisfied for the class $\bar{\mathcal{A}}$ and that, if (B2) can be verified, the optimal minimal values of the cost functional $J_\mu(u)$ over $\bar{\mathcal{A}}$ and \mathcal{A} coincide. Note, furthermore, that Proposition 1.1 and its Corollary 1.1, below, provide sufficient conditions for (B3) to hold. It is easily seen that the condition of Corollary 1.1 holds for Example (EX), provided that $g(x, u)$ is bounded and $f(x, u)$ satisfies, for all $x \in X$, the growth condition

$$(1.14) \quad \sup_{v \in U} \|f(x, v)\|_e^2 \leq \gamma(\|x\|_e^2 + K),$$

where $\|\cdot\|$ denotes the Euclidean norm, $\gamma \in (0, 1)$, and $0 < K < \infty$. Adding to these conditions the requirement that there exists $a > 0$ such that for all $x \in \mathcal{R}^n$, $v \in U$, $z \in \mathcal{R}^n$,

$$(1.15) \quad (g(x, v)z, g(x, v)z) \geq a\|z\|^2,$$

it is easily seen that in (EX) we have, for sufficiently large m ,

$$(1.16) \quad P\{\|x_1\| \geq m\} \geq P\left\{\|w_0\| \geq \frac{3m}{\sqrt{a}}\right\} = \alpha > 0,$$

so that (EX) also satisfies (B4). Assuming finally that we add to the state equation (1.6) an observation equation with $h(\cdot)$ satisfying (B1), we have in (EX) a rather general example for which (A1)–(A4) and (B1)–(B4) are all satisfied, provided that the admissible controls are restricted to $\bar{\mathcal{A}}$. Concerning assumption (B4), note finally that it represents a nondegeneracy condition for the state process that is always satisfied for state evolution models with nondegenerate additive Gaussian noise.

To study the case where E is compact, we assume the following assumptions in addition to (A1)–(A4).

(C1) There exists $j \in \{1, 2, \dots, d\}$ and $\hat{x} \in E$ such that either

$$(1.17) \quad h^j(\hat{x}) > h^j(x) \quad \text{for any } x \in E, x \neq \hat{x}$$

or

$$(1.18) \quad h^j(\hat{x}) < h^j(x) \quad \text{for any } x \in E, x \neq \hat{x}.$$

also require the following assumption.

(C2) There exists n_0 and a sequence of positive α_n such that for all $n \geq n_0$ we have

$$\inf_{v \in U} \inf_{x \in E} P^v(x, B_n) \geq \alpha_n$$

with $B_n = \{x \in E \mid \rho(x, \hat{x}) \leq n^{-1}\}$.

Contrary to the case where E is locally compact, but noncompact, provided it is possible to verify assumption (B2), we do not necessarily have to restrict the original

class \mathcal{A} of admissible control functions to the subclass $\tilde{\mathcal{A}}$, here we are forced to consider a subclass that we denote by \mathcal{A}^r . Letting $\psi(x) \in C(E)$ with values in $[0, 1]$ be any function such that

$$(1.19) \quad \hat{x} \notin \text{closure } \{x \mid \psi(x) > 0\},$$

the class \mathcal{A}^r is given by

$$(1.20) \quad \mathcal{A}^r = \{u \in \mathcal{A} \mid u(\nu) = \tilde{u}(\nu)r(\nu(\psi)) \text{ for } \tilde{u} \in \mathcal{A}\};$$

i.e., \mathcal{A}^r consists of control functions in \mathcal{A} that are equal to zero in a weak neighborhood of the Dirac measure $\delta_{\hat{x}}$.

Remark 1.2.3. Also for this case where E is compact, it follows from results to be obtained below that, given assumptions (A1)–(A4), (C1), (C2), as well as (B3), assumption (B2) is automatically satisfied for the class \mathcal{A}^r . In fact, by Theorem 2.3(ii), (iv) below, for each $u \in \mathcal{A}^r$ there exists a unique invariant measure Φ^u for the filtering process. Furthermore, by (A1), (A2) (see Proposition 1.2(ii) below), the filtering process $(\pi_i^{\mu, u})$ is, for each $u \in \mathcal{A}^r$, a Markov–Feller process; if then the initial law μ is such that (B3) holds, the ensuing tightness of the Cesaro averages together with the Feller property of $(\pi_i^{\mu, u})$, giving (1.10).

In the course of the paper, we make two further assumptions (A5) and (A6), which are also satisfied by Example (EX) under additional mild assumptions on the coefficients, and where we further require that the cost function $c(x, v)$ can be uniformly approximated by step functions in x .

Remark 1.2.4. The strongest of all our assumptions appears to be (A4). We can, however, provide an alternative approach to our problem, based on an extended notion of control that includes compulsory shifts on the state, depending on the value of the filter, and for which assumption (A4) is not required. Details are in [14].

We conclude this section with Proposition 1.1 and its Corollary 1.1, giving sufficient conditions for (B3) to hold. We use the symbol P_μ^u to denote the probability measure induced either by the signal process (x_i) starting from the initial measure μ , or by the corresponding filtering process $(\pi_i^{\mu, u})$.

PROPOSITION 1.1. *Suppose that for any $\varepsilon > 0$ there exists an increasing sequence of compact sets $L_k \subset E$ such that for any $u \in \mathcal{A}$*

$$(1.21) \quad \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} \sum_{k=1}^{\infty} 2^k P_\mu^u \{x_i \in L_k^c\} \leq \varepsilon;$$

then the family $\{t^{-1} \sum_{i=0}^{t-1} (\Pi^u)^i(\mu, \cdot), u \in \mathcal{A}, t = 1, 2, \dots\}$ is tight.

Proof. Let $\Gamma(\varepsilon) = \{\nu \in \mathcal{P}(E) : \nu(L_k^c) \leq 2^{-k} \text{ for all } k = 1, 2, \dots\}$. Clearly, $\Gamma(\varepsilon)$ is a compact subset of $\mathcal{P}(E)$. Moreover,

$$(1.22) \quad \begin{aligned} (\Pi^u)^i(\mu, \Gamma(\varepsilon)) &= P_\mu^u \{\pi_i^{\mu, u} \in \Gamma(\varepsilon)\} = P_\mu^u \{\pi_i^{\mu, u}(L_k^c) \leq 2^{-k} \text{ for all } k\} \\ &= 1 - P_\mu^u \{\pi_i^{\mu, u}(L_k^c) > 2^{-k} \text{ for some } k\} \\ &\geq 1 - \sum_{k=1}^{\infty} 2^k P_\mu^u \{x_i \in L_k^c\}. \end{aligned}$$

Therefore, by (1.21)

$$(1.23) \quad \liminf_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} (\Pi^u)^i(\mu, \Gamma(\varepsilon)) \geq 1 - \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} \sum_{k=1}^{\infty} 2^k P_\mu^u \{x_i \in L_k^c\} \geq 1 - \varepsilon,$$

from which our claim follows.

COROLLARY 1.1. Assume that $E = \mathcal{R}^n$ with Euclidean norm $\|\cdot\|_e$ and let

$$(1.24) \quad b = \sup_{u \in \mathcal{A}} \sup_i E_\mu^u \{\|x_i\|_e\} < \infty;$$

then (1.21) holds.

Proof. We define $L_k = \{x: \|x\|_e \leq 2^{2k} \varepsilon^{-1} b\}$. Then by Chebyshev's inequality

$$P_\mu^u \{\|x_i\|_e > 2^{2k} \varepsilon^{-1} b\} \leq 2^{-2k} \varepsilon b^{-1} E_\mu^u \{\|x_i\|_e\} \leq \varepsilon 2^{-2k},$$

and (1.21) is obviously satisfied.

1.3. Preliminary results. Consider the process (\bar{L}_n) defined as follows:

$$(1.25) \quad \bar{L}_n = \prod_{i=1}^n \exp \left[-(w_i, h(x_i)) - \frac{1}{2} (h(x_i), h(x_i)) \right]; \quad \bar{L}_0 = 1$$

with (\cdot, \cdot) standing for the Euclidean scalar product in \mathcal{R}^d . Given the increasing family of σ -algebras

$$(1.26) \quad (XW)^n := \sigma\{x_0, \dots, x_n; w_1, \dots, w_n\},$$

the process (\bar{L}_n) becomes a $((XW)^n, P)$ -martingale. Therefore we can define a new probability measure P^0 by

$$(1.27) \quad P_{(XW)^n}^0 = \bar{L}_n P_{(XW)^n} \quad \text{for } n = 1, 2, \dots$$

The following lemma is known and can be easily proved (see, e.g., [13]).

LEMMA 1.1. Under P^0 , the y_i are independently and identically distributed standard Gaussian, independent of the process (x_i) , and (x_i) has the same initial law μ and the same transition kernel $P^{u(\pi_i^{\mu,n})}(x_i, d\eta)$ as under P . Moreover,

$$(1.28) \quad P_{(XW)^n} = L_n P_{(XW)^n}^0$$

with

$$(1.29) \quad L_n = (\bar{L}_n)^{-1} = \prod_{i=1}^n \exp \left[(y_i, h(x_i)) - \frac{1}{2} (h(x_i), h(x_i)) \right], \quad L_0 = 1.$$

Using the conditional Bayes rule, usually called the Kallianpur–Striebel formula, we may now write

$$(1.30) \quad \pi_i^{\mu,u}(\phi) = \frac{E_\mu^0[L_i \phi(x_i) | Y^i]}{E_\mu^0[L_i | Y^i]} := \frac{\sigma_i^{\mu,u}(\phi)}{\sigma_i^{\mu,u}(1)} \quad P_\mu\text{-a.s.},$$

where, for $i = 1, 2, \dots$,

$$(1.31) \quad \begin{aligned} \sigma_i^{\mu,u}(\phi) &= E_\mu^0 \left[L_{i-1} E_\mu^0 \left\{ \exp \left[(y_i, h(x_i)) - \frac{1}{2} (h(x_i), h(x_i)) \right] \phi(x_i) \mid X^{i-1} Y^i \right\} \mid Y^i \right] \\ &= E_\mu^0 \left[L_{i-1} \int_E \exp \left[(y_i, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \phi(z) P^{u(\pi_{i-1}^{\mu,u})}(x_{i-1}, dz) \mid Y^i \right] \\ &= \int_E \int_E \exp \left[(y_i, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \phi(z) P^{u(\pi_{i-1}^{\mu,u})}(x, dz) \sigma_{i-1}^{\mu,u}(dx) \end{aligned}$$

with $\sigma_0^\mu(\cdot) = \mu(\cdot)$, and where E_μ^0 denotes expectation under the measure P^0 , assuming that (x_i) has initial law μ . By (1.30) it follows that the filtering process $(\pi_i^{\mu,u})$, defined in (1.2), can be obtained recursively as

$$\begin{aligned}
 \pi_0^{\mu,u}(\phi) &= \mu(\phi), \\
 \pi_i^{\mu,u}(\phi) &= \int_E \int_E \exp \left[(y_i, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \\
 &\quad \cdot \phi(z) P^{u(\pi_{i-1}^{\mu,u})}(x, dz) \pi_{i-1}^{\mu,u}(dx) \\
 &\quad \cdot \left(\int_E \int_E \exp \left[(y_i, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \right. \\
 &\quad \cdot P^{u(\pi_{i-1}^{\mu,u})}(x, dz) \pi_{i-1}^{\mu,u}(dx) \Big)^{-1} \\
 &:= M^u(y_i, \pi_{i-1}^{\mu,u})(\phi),
 \end{aligned}
 \tag{1.32}$$

where we implicitly define the operator M^u . The following proposition can now be proved (see [13]).

PROPOSITION 1.2. *Let $u \in \mathcal{A}$ be given and assumptions (A1), (A2) hold. Then*

- (i) *The mapping $M^u(y, \nu): \mathcal{R}^d \times \mathcal{P}(E) \rightarrow \mathcal{P}(E)$ is continuous;*
- (ii) *The process $(\pi_i^{\mu,u})$ is, for the filtration (Y^i) and under the probability measure P , a Feller–Markov process on $\mathcal{P}(E)$ with transition kernel $\Pi^u(\mu, \Lambda)$ given by*

$$\begin{aligned}
 \Pi^u(\mu, \Lambda) &= (2\pi)^{-(d/2)} \\
 &\quad \cdot \int_E \int_{\mathcal{R}^d} \left(\int_E \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] \right. \\
 &\quad \cdot P^{u(\mu)}(x, d\eta) 1_\Lambda(M^u(y, \mu)) \Big) dy \mu(dx)
 \end{aligned}
 \tag{1.33}$$

for $\Lambda \in \mathcal{B}(\mathcal{P}(E))$, the σ -field of Borel subsets of $\mathcal{P}(E)$.

To have a Feller property for $(\pi_i^{\mu,u})$, i.e., to know that Π^u transforms $C(\mathcal{P}(E))$ into itself, is very important for various reasons. We can then show, for example, that the Cesaro averages $n^{-1} \sum_{k=0}^{n-1} (\Pi^u)^k$, provided that they are tight and converge weakly to an invariant measure Φ^u of Π^u . Consequently, the functional (1.4) may be rewritten as

$$\begin{aligned}
 J_\mu(u) &= \limsup_{n \rightarrow \infty} n^{-1} \sum_{i=0}^{n-1} E_\mu \left\{ \int_E c(z, u(\pi_i^{\mu,u})) \pi_i^{\mu,u}(dz) \right\} \\
 &= \int_{\mathcal{P}(E)} \int_E c(x, u(\nu)) \nu(dx) \Phi^u(d\nu)
 \end{aligned}
 \tag{1.34}$$

for Φ^u -almost all $\mu \in \mathcal{P}(E)$.

2. Fundamental results.

2.1. Convergence of invariant measures. In this section, we prove Theorems 2.1 and 2.2 on convergence of invariant measures, where in the first theorem the control functions belong to the class \mathcal{A} , while in the second we restrict them to the classes $\bar{\mathcal{A}}$ or \mathcal{A}^r .

For this purpose, given $v \in U$, consider the transition kernel $P^v(x, \cdot)$ governing the given Markov process (x_i) on E and let $P_m^v(x, \cdot)$ be a sequence of Markov transition kernels approximating $P^v(x, \cdot)$ in the sense that

$$\begin{aligned}
 &\text{if } U \ni v_m \rightarrow v, \text{ then } P_m^{v_m}(x, \cdot) \Rightarrow P^v(x, \cdot) \\
 &\text{uniformly in } x \text{ from compact subsets of } E.
 \end{aligned}
 \tag{2.1}$$

Furthermore, given $u \in \mathcal{A}$ (or $\bar{\mathcal{A}}$, \mathcal{A}^r), let u_m be a sequence of Borel maps from $\mathcal{P}(E)$

into U such that

$$(2.2) \quad u_m(\nu) \rightarrow u(\nu) \text{ uniformly in } \nu \text{ from compact subsets of } \mathcal{P}(E).$$

Finally, let $h_m(x)$ be a sequence of bounded Borel maps approximating $h(x)$ in (1.1) in the sense that

$$(2.3) \quad \sup_{x \in E} |h_m(x) - h(x)| \rightarrow 0 \quad \text{for } m \rightarrow \infty.$$

Recall now that, given $u \in \mathcal{A}$ (or $\bar{\mathcal{A}}, \mathcal{A}^r$), $(\pi_i^{\mu, u})$ denotes the filtering process corresponding to the state process (x_i) with initial distribution μ , transition kernel $P^u(x, \cdot)$, and observations (1.1); note also that $P^u(x, \cdot)$ stands for the transition kernel that, in the generic period i , is given by $P^{v_i}(x, \cdot)$ with $v_i = u(\pi_i^{\mu, u})$.

Analogously, given $u \in \mathcal{A}$ (or $\bar{\mathcal{A}}, \mathcal{A}^r$) and $P_m^v(x, \cdot)$ as well as $u_m(\nu)$ according to (2.1) and (2.2), respectively, let (π_i^{m, μ, u_m}) denote the filtering process corresponding to the state process (x_i^m) with initial distribution μ , transition kernel $P_m^{u_m}(x_i^m, \cdot)$, and observations $y_i = h_m(x_i^m) + w_i$; again, $P_m^{u_m}(x, \cdot)$ stands for the transition kernel that, in the generic period i , is given by $P_m^{v_i}(x_i^m, \cdot)$ with $v_i = u_m(\pi_i^{m, \mu, u_m})$.

Below, we use the following additional notation: By analogy to $M^u(y, \nu)$ and $\Pi^u(\mu, \cdot)$ in (1.32), (1.33), we let

$$(2.4) \quad M_m^{u_m}(y, \nu)(\phi) := \int_E \int_E \exp \left[(y, h_m(z)) - \frac{1}{2} (h_m(z), h_m(z)) \right] \phi(z) P_m^{u_m(\nu)}(x, dz) \nu(dx) \\ \cdot \left(\int_E \int_E \exp \left[(y, h_m(z)) - \frac{1}{2} (h_m(z), h_m(z)) \right] P_m^{u_m(\nu)}(x, dz) \nu(dz) \right)^{-1}$$

for a bounded Borel function ϕ on E and use $\Pi_m^{u_m}(\mu, \cdot)$ to denote the transition kernel of the Markov process (π_i^{m, μ, u_m}) . Furthermore, by analogy to previous usage, we let $\Pi^u(\mu, F)$ and $\Pi_m^{u_m}(\mu, F)$ denote the integrals of a bounded Borel function $F(\nu) \in \mathcal{B}(\mathcal{P}(E))$ with respect to the measures $\Pi^u(\mu, d\nu)$ and $\Pi_m^{u_m}(\mu, d\nu)$, respectively.

We are now ready to state the first theorem.

THEOREM 2.1. *Given $u \in \mathcal{A}$, let (2.1)–(2.3) hold, as well as*

- (i) *Assumptions (A1), (A2);*
 - (ii) *$u \in \mathcal{A}$ is such that the corresponding filtering process $(\pi_i^{\mu, u})$ admits a unique invariant measure Φ^u ;*
 - (iii) *For each u_m satisfying (2.2), there exists an invariant measure $\Phi_m^{u_m}$ for the filtering process (π_i^{m, μ, u_m}) ;*
 - (iv) *The family $\{\Phi_m^{u_m}, m = 1, 2, \dots\}$ is tight.*
- Then $\Phi_m^{u_m} \Rightarrow \Phi^u$ for $m \rightarrow \infty$.*

THEOREM 2.2. *Let $u \in \bar{\mathcal{A}}$ or \mathcal{A}^r be given and (2.1)–(2.3) hold. Assume (A1)–(A4) as well as*

- (i) *For $u \in \bar{\mathcal{A}}$, assume furthermore (B1), (B4), with (B4) holding uniformly in m also for the sequences of kernels $P_m^v(x, \cdot)$ in (2.1);*
- (ii) *For $u \in \mathcal{A}^r$, assume furthermore (C1), (C2), with (C2) holding uniformly in m also for $P_m^v(x, \cdot)$.*

Then there exist unique invariant measures $\Phi_m^{u_m}, \Phi^u$ corresponding to the filtering processes $(\pi_i^{m, \mu, u_m}), (\pi_i^{\mu, u})$, respectively, and

$$\Phi_m^{u_m} \Rightarrow \Phi^u \quad \text{for } m \rightarrow \infty.$$

Furthermore, for any $F \in \mathcal{B}(\mathcal{P}(E))$, $\nu \in \mathcal{P}(E)$, any $u \in \bar{\mathcal{A}}$ or $u \in \mathcal{A}'$ satisfying the conditions in (i) and (ii), respectively, and for sufficiently large m ,

$$(2.5) \quad \lim_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} (\Pi_m^{u_m}(\nu, F))^i = \int_{\mathcal{P}(E)} F(\nu) \Phi_m^{u_m}(d\nu),$$

$$(2.6) \quad \lim_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} (\Pi^u(\nu, F))^i = \int_{\mathcal{P}(E)} F(\nu) \Phi^u(d\nu).$$

For the proof of the above theorems, the following proposition and its corollary will be crucial.

PROPOSITION 2.1. *Let (2.1)–(2.3) hold and assume (A1), (A2). If $F_m \in \mathcal{B}(\mathcal{P}(E))$ is a sequence of Borel functions with uniform bound b_F , i.e.,*

$$(2.7) \quad \|F_m\| \leq b_F$$

and satisfies

$$(2.8) \quad F_m \rightarrow F \in C(\mathcal{P}(E))$$

uniformly on compact subsets of $\mathcal{P}(E)$, then

$$(2.9) \quad \Pi_m^{u_m}(\nu, F_m) \rightarrow \Pi^u(\nu, F)$$

uniformly in ν from compact subsets of $\mathcal{P}(E)$.

Proof. See Appendix A.1.

Corollary 2.1 follows immediately.

COROLLARY 2.1. *Under the assumptions of Proposition 2.1, if $\nu_m \Rightarrow \nu$, then*

$$\Pi_m^{u_m}(\nu_m, F_m) \rightarrow \Pi^u(\nu, F) \quad \text{for } m \rightarrow \infty.$$

Proof of Theorem 2.1. By the tightness assumption (iv) of Theorem 2.1, we have that the family $\{\Phi_m^{u_m}, m = 1, 2, \dots\}$ has a compact closure [4, Thm. 6.1] and there exists a measure $\Phi \in \mathcal{P}(\mathcal{P}(E))$ and a subsequence of $\{m\}$, for simplicity also denoted by $\{m\}$, such that $\Phi_m^{u_m} \Rightarrow \Phi$. Then, for $F \in C(\mathcal{P}(E))$,

$$(2.10) \quad \begin{aligned} |\Phi(F) - \Phi(\Pi^u F)| &\leq |\Phi(F) - \Phi_m^{u_m}(F)| + |\Phi_m^{u_m}(F) - \Phi_m^{u_m}(\Pi_m^{u_m} F)| \\ &\quad + |\Phi_m^{u_m}(\Pi_m^{u_m} F) - \Phi_m^{u_m}(\Pi^u F)| + |\Phi_m^{u_m}(\Pi^u F) - \Phi(\Pi^u F)|. \end{aligned}$$

Since $\{\Phi_m^{u_m}, m = 1, 2, \dots\}$ is tight, for each $\varepsilon > 0$ there is a compact set $\Gamma \subset \mathcal{P}(E)$ such that

$$\Phi_m^{u_m}(\Gamma) \geq 1 - \varepsilon \quad \text{for } m = 1, 2, \dots$$

Hence

$$(2.11) \quad |\Phi_m^{u_m}(\Pi_m^{u_m} F) - \Phi_m^{u_m}(\Pi^u F)| \leq 2\|F\|\varepsilon + \sup_{\nu \in \Gamma} |\Pi_m^{u_m} F(\nu) - \Pi^u F(\nu)|.$$

By (A1), (A2), and Proposition 1.2, we know that $\Pi^u F \in C(\mathcal{P}(E))$. Therefore, letting $m \rightarrow \infty$ in (2.10), from Proposition 2.1 and (2.11), we obtain

$$(2.12) \quad |\Phi(F) - \Phi(\Pi^u F)| \leq 2\|F\|\varepsilon.$$

Thus, since ε can be chosen arbitrarily small, we obtain $\Phi(F) = \Phi(\Pi^u F)$ for any $F \in \mathcal{P}(E)$, and, by the uniqueness of the invariant measure of $(\pi_t^{\mu, u})$, we have $\Phi = \Phi^u$. \square

For the proof of Theorem 2.2, we need additional notation and further preliminary results. Concerning the notation, let $\bar{\psi}_n(x) \in C(E)$ with values in $[0, 1]$ be the same as

$\psi_n(x)$ in (1.12) if (B1), (B2) hold, while in case of (C1), (C2) it is given by

$$(2.13) \quad \bar{\psi}_n(x) = \begin{cases} 0 & \text{for } x \in \{z \in E \mid \rho(z, \hat{x}) \leq 1/n\}, \\ 1 & \text{for } x \in \{z \in E \mid \rho(z, \hat{x}) \geq 2/n\} \end{cases}$$

and is furthermore Lipschitz with Lipschitz constant n .

Consider now $u \in \mathcal{A}$; then (see (1.13)), for some n , $u \in \bar{\mathcal{A}}_n$. On the other hand, if $u \in \mathcal{A}'$, then (see (1.20)), for some n , $\{z \in E \mid \rho(z, \hat{x}) \leq 2/n\} \subset \{x \mid \psi(x) = 0\}$. Whether $u \in \bar{\mathcal{A}}$ or $u \in \mathcal{A}'$, let \tilde{n} be a positive integer with the properties just mentioned and define

$$(2.14) \quad \Gamma = \{\nu \in \mathcal{P}(E), \nu(\bar{\psi}_{\tilde{n}}) < b\},$$

where b is as in the definition of the function $r(\cdot)$ in (1.11). Consider then the stopping times

$$\begin{aligned} \tau^{\mu, u} &= \inf \{i > 0 \mid \pi_i^{\mu, u} \in \Gamma\} = \inf \{i > 0 \mid \pi_i^{\mu, u}(\bar{\psi}_{\tilde{n}}) < b\}, \\ \bar{\tau}^{\mu, u} &= \inf \{i \geq 0 \mid \pi_i^{\mu, u} \in \Gamma\} \\ \tau_{(m)}^{\mu, u_m} &= \inf \{i > 0 \mid \pi_i^{m, \mu, u_m} \in \Gamma\}, \\ \bar{\tau}_{(m)}^{\mu, u_m} &= \inf \{i \geq 0 \mid \pi_i^{m, \mu, u_m} \in \Gamma\} \end{aligned}$$

as well as the sequences

$$\begin{aligned} \tau_1^{\mu, u} &= \tau^{\mu, u}, & \tau_{n+1}^{\mu, u} &= \tau_n^{\mu, u} + \tau^{\mu, u} \Theta_{\tau_n^{\mu, u}}, \\ \tau_{1(m)}^{\mu, u_m} &= \tau_{(m)}^{\mu, u_m}, & \tau_{n+1(m)}^{\mu, u_m} &= \tau_{n(m)}^{\mu, u_m} + \tau_{(m)}^{\mu, u_m} \Theta_{\tau_{n(m)}^{\mu, u_m}}, \end{aligned}$$

where Θ_{τ_n} stands for the Markov shift operator of the corresponding filtering process. Finally, for given $y \in \mathcal{R}^d$, $\mu \in \mathcal{P}(E)$, define the measures $R(y, \mu)$, $R_m(y, \mu) \in \mathcal{P}(E)$ by ($K \in \mathcal{B}(E)$) as follows:

$$(2.15) \quad R(y, \mu)(K) = \frac{\int_K \exp[(y, h(z)) - \frac{1}{2}(h(z), h(z))] \mu(dz)}{\int_E \exp[(y, h(z)) - \frac{1}{2}(h(z), h(z))] \mu(dz)},$$

$$(2.16) \quad R_m(y, \mu)(K) = \frac{\int_K \exp[(y, h_m(z)) - \frac{1}{2}(h_m(z), h_m(z))] \mu(dz)}{\int_E \exp[(y, h_m(z)) - \frac{1}{2}(h_m(z), h_m(z))] \mu(dz)},$$

so that, letting

$$(2.17) \quad \zeta^u(\mu)(\cdot) = \int_E P^{u(\mu)}(x, \cdot) \mu(dx), \quad \zeta_{m^m}^{u_m}(\mu)(\cdot) = \int_E P^{u_m(\mu)}(x, \cdot) \mu(dx),$$

we have (see (1.32), (2.4))

$$(2.18) \quad M^u(y, \mu) = R(y, \zeta^u(\mu)), \quad M_{m^m}^{u_m}(y, \mu) = R_m(y, \zeta_{m^m}^{u_m}(\mu)).$$

LEMMA 2.1. Under (B1), (B4) or (C1), (C2), for any $\gamma \in (0, 1)$, $M > 0$, $n = 1, 2, \dots$, and any $u \in \mathcal{A}$ we can find y_0, m_0 such that, if $(y^j$ is the j th component of $y \in \mathcal{R}^d)$

$$(2.19) \quad y^j > y_0, \quad |y^i| \leq M \quad \text{when } i \neq j,$$

we have for any $\nu \in \mathcal{P}(E)$ and $m > m_0$

$$(2.20) \quad M^u(y, \nu)(K_n^c) \geq \gamma, \quad M_{m^m}^{u_m}(y, \nu)(K_n^c) \geq \gamma,$$

$$(2.21) \quad M^u(y, \nu)(B_n) \geq \gamma, \quad M_{m^m}^{u_m}(y, \nu)(B_n) \geq \gamma,$$

according to whether (B1), (B4) or (C1), (C2) are satisfied and where K_n is as in (1.7), and B_n as in (C2).

Proof. See Appendix A.1.

From Lemma 2.1 and the above definitions, we immediately have the following corollary.

COROLLARY 2.2. Under (B1), (B4) or (C1), (C2),

$$\inf_{u \in \mathcal{A}} \inf_{\nu \in \mathcal{P}(E)} \Pi^u(\nu, \Gamma) \geq \beta > 0,$$

$$\inf_{u \in \mathcal{A}} \inf_{\nu \in \mathcal{P}(E)} \inf_{m \geq m_0} \Pi_{m^m}^u(\nu, \Gamma) \geq \beta > 0,$$

Consequently, the following holds.

COROLLARY 2.3. Always under (B1), (B4) or (C1), (C2), for $k = 1, 2, \dots$, we have

$$\sup_{u \in \mathcal{A}} \sup_{\nu \in \mathcal{P}(E)} E_{\nu}^u(\tau^{\nu, u})^k < \infty,$$

$$\sup_{u \in \mathcal{A}} \sup_{\nu \in \mathcal{P}(E)} \sup_{m \geq m_0} E_{\nu}^{u_m}(\tau_{(m)}^{\nu, u_m})^k < \infty,$$

which implies the same statement also for $\bar{\tau}^{\nu, u}$ and $\bar{\tau}_{(m)}^{\nu, u_m}$.

We also have the following lemmas.

LEMMA 2.2. Under the assumptions of Theorem 2.2, the processes $(\pi_{\tau_n}^{\nu, u})$ and $(\pi_{\tau_{n(m)}}^{m, \mu, u_m})$ —where for simplicity of notation we identify τ_n with $\tau_n^{\nu, u}$ and $\tau_{n(m)}$ with $\tau_{n(m)}^{\nu, u}$ —have, for m greater than the m_0 in Lemma 2.1 and for $\nu \in \Gamma$, unique invariant measures $I^u(\cdot)$ and $I_m^{u_m}(\cdot)$, respectively. More precisely, we have for $B \in \mathcal{B}(\mathcal{P}(E))$

$$\begin{aligned} I^u(B) &= E_{\nu}^u\{\pi_{\tau_1}^{\nu, u} \in B\} \\ (2.22) \quad &= \int_E (2\pi)^{-d/2} \int_{\mathcal{R}^d} E_{R(y, \eta)}^u\{\pi_{\bar{\tau}}^{R, u} \in B\} \\ &\quad \cdot \exp\left[-\frac{1}{2}(y - h(z), y - h(z))\right] dy \eta(dz), \end{aligned}$$

$$\begin{aligned} I_m^{u_m}(B) &= E_{\nu}^{u_m}\{\pi_{\tau_{1(m)}}^{m, \nu, u_m} \in B\} \\ (2.23) \quad &= \int_E (2\pi)^{-d/2} \int_{\mathcal{R}^d} E_{R_m(y, \eta)}^{u_m}\{\pi_{\bar{\tau}(m)}^{m, R, u_m} \in B\} \\ &\quad \cdot \exp\left[-\frac{1}{2}(y - h(z), y - h(z))\right] dy \eta(dz), \end{aligned}$$

where $\eta(\cdot)$ is the measure corresponding to (A4) and E_{ν}^u denotes expectation under the control function u and initial distribution $\nu \in \mathcal{P}(E)$.

Proof. The proof uses the definition of the set Γ in (2.14) and of the stopping times below (2.14) as well as the properties of the controls $u \in \bar{\mathcal{A}}$ or $u \in \mathcal{A}^r$ mentioned above (2.14). \square

LEMMA 2.3. Under the assumptions of Theorem 2.2, for $m > m_0$ there exist unique invariant measures $\Phi^u, \Phi_m^{u_m}$ of the processes $\{\pi_i^{\mu, u}\}$ and $\{\pi_i^{m, \mu, u_m}\}$, respectively. Letting $B \in \mathcal{B}(\mathcal{P}(E))$, they are given by

$$\begin{aligned} \Phi^u(B) &= \int_E (2\pi)^{-d/2} \int_{\mathcal{R}^d} E_{R(y, \eta)}^u \left\{ \sum_{i=0}^{\bar{\tau}} \chi_B(\pi_i^{R, u}) \right\} \\ (2.24) \quad &\cdot \exp\left[-\frac{1}{2}(y - h(z), y - h(z))\right] dy \eta(dz) \\ &\cdot \left(\int_E (2\pi)^{-d/2} \int_{\mathcal{R}^d} E_{R(y, \eta)}^u \{\bar{\tau} + 1\} \right. \\ &\quad \left. \cdot \exp\left[-\frac{1}{2}(y - h(z), y - h(z))\right] dy \eta(dz) \right)^{-1}, \end{aligned}$$

$$\begin{aligned}
\Phi_m^{u_m}(B) = & \int_E (2\pi)^{-d/2} \int_{\mathbb{R}^d} E_{R_m(y,\eta)}^{u_m} \left\{ \sum_{i=0}^{\bar{\tau}_{(m)}} \chi_B(\pi_i^{m,R_m,u_m}) \right\} \\
& \cdot \exp \left[-\frac{1}{2} (y - h(z), y - h(z)) \right] dy \eta(dz) \\
(2.25) \quad & \cdot \left(\int_E (2\pi)^{-d/2} \int_{\mathbb{R}^d} E_{R_m(y,\eta)}^{u_m} \{ \bar{\tau}_{(m)} + 1 \} \right. \\
& \left. \cdot \exp \left[-\frac{1}{2} (y - h(z), y - h(z)) \right] dy \eta(dz) \right)^{-1}.
\end{aligned}$$

Proof. The proof uses the following intermediate result. Since by the previous lemma there exists a unique invariant measure I^u for $\pi_{\tau_n}^{\nu,u}$ with $\nu \in \Gamma$, by Corollary 2.3 there exists a unique invariant measure Φ^u also for $(\pi_i^{\mu,u})$ with $\nu \in \mathcal{P}(E)$ that is given by

$$\Phi^u(B) = \frac{\int_{\Gamma} E_{\nu}^u \{ \sum_{i=0}^{\tau-1} \chi_B(\pi_i^{R,u}) \} I^u(d\nu)}{\int_{\Gamma} E_{\nu}^u \{ \tau \} I^u(d\nu)}$$

and analogously for the process (π_i^{m,μ,u_m}) . From here, we then proceed by analogy to the proof of the previous lemma, exploiting the definition of the set Γ and the structure of the controls $u \in \tilde{\mathcal{A}}$ or $u \in \mathcal{A}^r$. \square

At this point, to prove Theorem 2.2 it suffices to show the convergence of the right-hand side in (2.25) to the right-hand side in (2.24). By Corollary 2.3, this, in turn, is equivalent to showing that

$$\begin{aligned}
(2.26) \quad & E_{R_m(y,\eta)}^{u_m} \{ \chi_{\Gamma^c}(\pi_1^{m,R_m,u_m}) \cdots \chi_{\Gamma^c}(\pi_{i-1}^{m,R_m,u_m}) F(\pi_i^{m,R_m,u_m}) \} \\
& \rightarrow E_{R(y,\eta)}^u \{ \chi_{\Gamma^c}(\pi_1^{R,u}) \cdots \chi_{\Gamma^c}(\pi_{i-1}^{R,u}) F(\pi_i^{R,u}) \}.
\end{aligned}$$

Because the function of $(\pi_1^{R,u}, \dots, \pi_i^{R,u})$ on the right-hand side of (2.26) is not continuous, we also need the following auxiliary result, whose proof is in Appendix A.1.

LEMMA 2.4. *Let $u \in \tilde{\mathcal{A}}(\mathcal{A}^r)$ and Z be a continuous \mathbb{R}^d -valued random variable; furthermore, let $\nu \in \mathcal{P}(E)$ be such that, if $\mathcal{O} \subset E$ is an open set, then $\nu(\mathcal{O}) > 0$. The following then hold:*

(i) *Under (B1), for any $\phi \in C(E)$ such that $\text{supp}(\phi) \subset K_n$ and any $b \in (0, 1)$, we have*

$$P\{M^u(Z, \nu)(\phi) = b\} = 0;$$

(ii) *Under (C1), for any $\phi \in C(E)$ such that $\text{supp}(\phi) \subset E \setminus B_n$ and any $b \in (0, 1)$, we have*

$$P\{M^u(Z, \nu)(\phi) = b\} = 0.$$

We finally also have Proposition 2.2, whose proof follows immediately from Lemma A.1.3 in Appendix A.1 and the fact that, for all $y \in \mathbb{R}^d$, $\nu \in \mathcal{P}(E)$,

$$(2.27) \quad R_m(y, \nu) \rightarrow R(y, \nu).$$

PROPOSITION 2.2. *Under the assumptions of Theorem 2.2, given a positive integer q and $F_1, \dots, F_q \in C(\mathcal{P}(E))$, we have, for all $(y, \nu) \in \mathbb{R}^d \times \mathcal{P}(E)$,*

$$\begin{aligned}
& E_{R_m(y,\eta)}^{u_m} \{ F_1(\pi_1^{m,R_m,u_m}) \cdots F_q(\pi_q^{m,R_m,u_m}) \} \\
& \rightarrow E_{R(y,\eta)}^u \{ F_1(\pi_1^{R,u}) \cdots F_q(\pi_q^{R,u}) \} \quad \text{for } m \rightarrow \infty.
\end{aligned}$$

We are now in a position to complete the following proof.

Proof of Theorem 2.2. The statement of Proposition 2.2 is equivalent to the weak convergence

$$(\pi_1^{m, R_m, u_m}, \dots, \pi_q^{m, R_m, u_m}) \Rightarrow (\pi_1^{R, u}, \dots, \pi_q^{R, u}).$$

By Lemma 2.4 we then obtain (2.26) from which we get the desired weak convergence of the invariant measure, noting that

$$\begin{aligned} E_{R_m(y, \eta)}^{u_m} \left\{ \sum_{i=0}^{\bar{\tau}(m)} \chi_B(\pi_i^{m, R_m, u_m}) \right\} \\ = \chi_B(R_m(y, \eta)) + \sum_{i=1}^{\infty} \chi_{\Gamma^c}(R_m(y, \eta)) \\ \cdot E_{R_m(y, \eta)}^{u_m} \{ \chi_{\Gamma^c}(\pi_1^{m, R_m, u_m}) \cdot \dots \cdot \chi_{\Gamma^c}(\pi_{i-1}^{m, R_m, u_m}) \chi_B(\pi_i^{m, R_m, u_m}) \}. \end{aligned}$$

The last statement of the theorem follows from the law of large numbers applied to the uniformly ergodic processes $(\pi_{\tau_n}^{v, u})$ and $(\pi_{\tau_n(m)}^{m, \mu, u_m})$ (see Lemma 2.2) and from Corollary 2.3 (details are similar to Proposition 3 in [7]). \square

2.2. Approximation of admissible control functions. For our approximation purpose, it is crucial to have not only a unique limiting invariant measure, but also control functions $u: \mathcal{P}(E) \rightarrow U$ of a possibly simple form. We therefore approximate the admissible control functions in the classes $\mathcal{A}, \bar{\mathcal{A}}, \mathcal{A}^r$ considered so far, by simpler control functions in corresponding families parametrized by two parameters $L > 0$ and $n = 1, 2, \dots$, in such a way that

(a) For all approximating control functions u , the controlled filtering process $(\pi_i^{\mu, u})$ admits a unique invariant measure Φ^u ;

(b) The infima of the objective function $J_\mu(u)$ over the classes $\mathcal{A}, \bar{\mathcal{A}}, \mathcal{A}^r$ and over the corresponding classes of approximating control functions can be made as close as possible by choosing the parameters L and n sufficiently large.

DEFINITION 2.1. If E is noncompact and (B1) holds, then, given $L > 0$ and integer $n > 0$, let

$$\mathcal{A}(L, n) := \{u \in \mathcal{A} \mid u(\nu) = \bar{u}(\nu(\phi_1), \dots, \nu(\phi_n))r(\nu(\psi_n))\},$$

where

- $\bar{u}: \mathcal{R}^n \rightarrow U$ is Lipschitz with Lipschitz constant L ,
- $\phi_1, \phi_2, \dots, \phi_n, \dots$ is a dense sequence in $C_0(E)$, the space of continuous functions vanishing at infinity, and
- $r(\cdot)$ and $\psi_n(\cdot)$ are fixed and given as in (1.11), (1.12).

Note that $\mathcal{A}(L, n)$ is a subclass not only of \mathcal{A} , but also of $\bar{\mathcal{A}}$.

DEFINITION 2.2. If (C1) holds, then, given $L > 0$ and integer $n > 0$, let

$$\mathcal{A}^r(L, n) := \{u \in \mathcal{A}^r \mid u(\nu) = \bar{u}(\nu(\phi_1), \dots, \nu(\phi_n))r(\nu(\psi \wedge \psi_n))\},$$

where \bar{u} and $r(\cdot)$ are, as in Definition 2.1,

- $\phi_1, \phi_2, \dots, \phi_n, \dots$ is a dense sequence in $C(E)$ or $C_0(E)$ according to whether E is compact or only locally compact,
- $\psi(\cdot)$ is any function as in (1.19) with the additional requirement of being Lipschitz with constant n , and
- $\psi_n(\cdot)$ are given Lipschitz functions with constant n , taking values in $[0, 1]$, and such that

$$\psi_n(x) = \begin{cases} 0 & \text{for } x \in \{z \in E: \rho(z, \hat{x}) < 1/n\} \cup E \setminus K_n, \\ 1 & \text{for } x \in K_{n-1} \setminus \{z \in E: \rho(z, \hat{x}) < 2/n\}, \end{cases}$$

where K_n is as in (1.7).

We now state Theorem 2.3, which will not only be of interest in itself, namely, to justify the choice of the classes $\mathcal{A}(L, n)$, $\mathcal{A}^r(L, n)$, but its setup will be useful also in various situations in the rest of the paper, where it is important to have a unique limiting invariant measure (see, e.g., Proposition 3.2 and Theorem 3.1 below).

THEOREM 2.3. *Assume (A1)–(A4). Then*

(i) *If (B1)–(B3) are satisfied, for each $\mu \in \mathcal{P}(E)$ for which (B3) holds, we have*

$$(2.28) \quad \lim_{L \rightarrow \infty, n \rightarrow \infty} \inf_{u \in \mathcal{A}(L, n)} J_\mu(u) = \inf_{u \in \mathcal{A}} J_\mu(u);$$

(ii) *If (B3), (C1) are satisfied, for each $\mu \in \mathcal{P}(E)$ for which (B3) holds, we have*

$$(2.29) \quad \lim_{L \rightarrow \infty, n \rightarrow \infty} \inf_{u \in \mathcal{A}^r(L, n)} J_\mu(u) = \inf_{u \in \mathcal{A}^r} J_\mu(u);$$

(iii) *If (B1), (B4) are satisfied, for any $\mu \in \mathcal{P}(E)$, we have*

$$(2.3) \quad \lim_{L \rightarrow \infty, n \rightarrow \infty} \inf_{u \in \mathcal{A}(L, n)} J_\mu(u) = \inf_{u \in \mathcal{A}} J_\mu(u);$$

(iv) *If (C1), (C2) are satisfied, for any $\mu \in \mathcal{P}(E)$, we have*

$$(2.31) \quad \lim_{L \rightarrow \infty, n \rightarrow \infty} \inf_{u \in \mathcal{A}^r(L, n)} J_\mu(u) = \inf_{u \in \mathcal{A}^r} J_\mu(u).$$

Moreover, the controlled filtering process $(\pi_i^{\mu, u})$ admits a unique invariant measure Φ^u in the following situations:

For $u \in \mathcal{A}(L, n)$ under the assumptions of case (i);

For $u \in \mathcal{A}^r$ under the assumptions of either case (ii) or (iv);

For $u \in \tilde{\mathcal{A}}$ under the assumptions of case (iii).

Proof. By the very definition of compact sets in $\mathcal{P}(E)$, for each compact set $\Lambda \in \mathcal{P}(E)$, and (ψ_n) as in (1.12), we have $r(\nu(\psi_n)) = 1$ for $\nu \in \Lambda$ and n sufficiently large. Moreover, for any $\psi \in C(E)$, for which $0 \leq \psi \leq 1$, $\hat{x} \notin \text{closure } \{x: \psi(x) > 0\}$ there is a sequence (g_n) of Lipschitz functions taking values in $[0, 1]$ and having Lipschitz constant n , such that, with (ψ_n) as in Definition 2.2, we have by the Stone–Weierstrass theorem (see, e.g., [11, Thm. 9.28])

$$(2.32) \quad \lim_{n \rightarrow \infty} \min \{\psi_n(x), g_n(x)\} = \psi(x)$$

uniformly on compact subsets of E , from which we obtain the convergence $r(\nu(\psi_n \wedge g_n)) \rightarrow r(\nu(\psi))$, for $n \rightarrow \infty$, uniformly in ν from compact subsets Λ of $\mathcal{P}(E)$. Furthermore, by a suitable version of the Stone–Weierstrass theorem (see [16, App.]), each $u \in C(\mathcal{P}(E), U)$ can be approximated uniformly on compact sets with the use of functions $\hat{u}(\nu) = \bar{u}(\nu(\phi_1), \dots, \nu(\phi_n))$, $n = 1, 2, \dots$, $\bar{u} \in C(\mathcal{R}^n, U)$. Therefore, each $u \in \mathcal{A}$, respectively \mathcal{A}^r , can be approximated uniformly on compact sets with the use of functions of the type $u(\nu)$ from the respective classes $\mathcal{A}(L, n)$, $\mathcal{A}^r(l, n)$, with L, n sufficiently large.

Starting with case (i), note that, for each $u_\varepsilon \in \mathcal{A}$ satisfying (B2), there exists a sequence L_m with $L_m \leq L_{m+1} \rightarrow \infty$, and a sequence $u_m \in \mathcal{A}(L_m, m)$, such that $u_m(\nu) \rightarrow u_\varepsilon(\nu)$ uniformly on compact subsets of $\mathcal{P}(E)$. Since by Proposition 1.2, under (A1)–(A2) each process (π_i^{μ, u_m}) is Feller, and by (B3) the family

$$\left\{ \frac{1}{t} \sum_{i=0}^{t-1} (\Pi^{u_m})^i(\mu, \cdot), t = 1, 2, \dots \right\}$$

is tight, any weak limit Φ^{u_m} of

$$\frac{1}{t} \sum_{i=0}^{t-1} (\Pi^{u_m})^i(\mu, \cdot) \quad \text{with } t \rightarrow \infty$$

is an invariant measure of Π^{u_m} . Moreover, by (B3) again, the family $\{\Phi^{u_m}, m = 1, 2, \dots\}$ is tight. Therefore, recalling that by (B2) we have a unique invariant measure Φ^{u^*} , the assumptions of Theorem 2.1. are satisfied, and we have $\Phi^{u_m} \Rightarrow \Phi^{u^*}$ for $m \rightarrow \infty$. Consequently, (2.28) holds.

To prove (2.29), let us for a moment assume the uniqueness of an invariant measure Φ^u of the controlled filtering process $\pi^{\mu, u}$ for $u \in \mathcal{A}^r$. Given this fact and a sequence $u_m \in \mathcal{A}^r(L_m, m)$ ($L_m, m \rightarrow \infty$) such that $u_m(\nu) \rightarrow u(\nu)$ uniformly on compact subsets of $\mathcal{P}(E)$, since by (B3) the family $\{\Phi^{u_m}, m = 1, 2, \dots\}$ is tight, all the assumptions of Theorem 2.1 are again fulfilled. Therefore $\Phi^{u_m} \Rightarrow \Phi^u$ for $m \rightarrow \infty$, and, taking into account that for each $\mu \in \mathcal{P}(E)$ for which (B3) holds and $u \in \mathcal{A}^r$

$$\frac{1}{t} \sum_{i=0}^{t-1} (\Pi^u)^i(\mu, \cdot) \Rightarrow \Phi^u, \quad \text{letting } t \rightarrow \infty,$$

we obtain (2.29).

The proofs for the cases (iii) and (iv) are based on Theorem 2.2 and follow almost immediately from (2.5) and (2.6).

It remains to show the second assertion of the theorem. By results of the proof of Theorem 1.21 and the remark following Theorem 1.4 in [15], the existence of a unique invariant set, i.e., the fact that there are no two disjoint invariant sets, ensures the existence of a unique invariant measure. Therefore we prove that there exists a unique invariant set for the process $\pi_i^{\mu, u}$, $\mu \in \mathcal{P}(E)$.

Let $u \in \mathcal{A}(L, n)$. For $i = 0, 1, 2, \dots$, define

$$(2.33) \quad \pi_{i+1}^{\mu, u}(\phi, y) = \int_E \exp \left[(y, h(\eta)) - \frac{1}{2} (h(\eta), h(\eta)) \right] \phi(\eta) \hat{\pi}_i^{\mu, u}(d\eta) \cdot \left(\int_E \exp \left[(y, h(\eta)) - \frac{1}{2} (h(\eta), h(\eta)) \right] \hat{\pi}_i^{\mu, u}(d\eta) \right)^{-1}$$

with

$$(2.34) \quad \hat{\pi}_i^{\mu, u}(d\eta) = \int_E P^{u(\pi_i^{\mu, u})}(x, d\eta) \pi_i^{\mu, u}(dx).$$

Clearly, $\pi_{i+1}^{\mu, u}(\phi) = \pi_{i+1}^{\mu, u}(\phi, y_{i+1})$. Assume that (1.8) is satisfied. Then

$$(2.35) \quad \pi_{i+1}^{\mu, u}(\psi_n, y) \rightarrow 0 \text{ for the } j\text{th coordinate } y^j \text{ of } y \text{ tending to } \infty.$$

In fact, let $\|h^j\|_n = \sup_{x \in K_n} |h^j(x)|$. By (1.8), $\|h^j\| > \|h^j\|_n$ and

$$(2.36) \quad \pi_{i+1}^{\mu, u}(\psi_n, y) = \frac{\int_{K_n} \psi_n(z) \exp[(y, h(z)) - \|h^j\|_n - \frac{1}{2}(h(z), h(z))] \hat{\pi}_i^{\mu, u}(dz)}{\int_E \exp[(y, h(z)) - \|h^j\|_n - \frac{1}{2}(h(z), h(z))] \hat{\pi}_i^{\mu, u}(dz)}.$$

From (A3) and (2.34), $\hat{\pi}_i^{\mu, u}(\mathcal{O}) > 0$ for some open set \mathcal{O} , implying (2.35).

Now, by (2.35) there are $d_j > 0$, $d > 0$, such that, for $y^j > d_j$ and $|y^k| \leq d$ with $k \neq j$, we have $\pi_{i+1}^{\mu, u}(\psi_n, y) < b$ with b as in (1.11). Therefore $r(\pi_{i+1}^{\mu, u}(\psi_n)) = 0$, and, consequently, $u(\pi_{i+1}^{\mu, u}) = 0$, all with a positive probability. Thus, on account of assumption (A4),

$$(2.37) \quad \pi_{i+2}^{\mu, u}(\phi, y_{i+2}) = \frac{\int_E \exp[(y_{i+2}, h(z)) - \frac{1}{2}(h(z), h(z))] \phi(z) \eta(dz)}{\int_E \exp[(y_{i+2}, h(z)) - \frac{1}{2}(h(z), h(z))] \eta(dz)}$$

with positive probability. Since for any other $\nu \in \mathcal{P}(E)$, again with positive probability, $\pi_{i+2}^{\nu, u}(\phi, y_{i+2})$ is equal to the right-hand side of (2.37), there is a unique invariant set for the filtering process $(\pi_i^{\mu, u})$. The proof of the uniqueness of the invariant measure Φ^u for $u \in \mathcal{A}(L, n)$ and under (1.8) is thus completed. The case where $u \in \mathcal{A}(L, n)$ and

under (1.9) can be shown analogously. The proof of the uniqueness of Φ^u for $u \in \mathcal{A}'$ and under (C1) consists of similar steps: namely, we show that $\pi_{i+1}^{\mu,u}(\psi, y) \rightarrow 0$ for $y^j \rightarrow \infty$ under (1.17) or for $y^j \rightarrow -\infty$ under (1.18), which by (A3) and (A4) implies that $r(\pi_{i+1}^{\mu,u}(\psi)) = 0$, and, consequently, $u(\pi_{i+1}^{\mu,u}) = 0$, with a positive probability. Therefore, for any $\nu \in \mathcal{P}(E)$, $\pi_{i+2}^{\mu,u}$ is equal with positive probability to the right-hand side of (2.37), which implies the uniqueness of the invariant set and, consequently, of the invariant measure. The cases of $u \in \mathcal{A}'$ under (C1), (C2) and of $u \in \bar{\mathcal{A}}$ under (B1), (B4) follow immediately from Theorem 2.2 and the definitions of the approximating control functions in the classes $\mathcal{A}(L, n)$, $\mathcal{A}'(L, n)$. \square

COROLLARY 2.4. *Under (A1)–(A4), (B1), (B4) for the class $\bar{\mathcal{A}}$ of admissible control functions defined in (1.13), we have that assumption (B2) is satisfied. If (B2) can be verified independently, we also have for any $\mu \in \mathcal{P}(E)$*

$$\inf_{u \in \bar{\mathcal{A}}} J_\mu(u) = \inf_{u \in \mathcal{A}} J_\mu(u).$$

3. Construction of a nearly optimal control function.

3.1. Approximation of the state process. So far, the admissible control functions have been reduced to those that are elements of $\mathcal{A}(L, n)$ or $\mathcal{A}'(L, n)$, but it will be difficult to determine an optimal or even only ε -optimal control for $J_\mu(u)$ in the classes $\mathcal{A}(L, n)$ or $\mathcal{A}'(L, n)$. We therefore need to further approximate the problem at hand; the first step in this direction is a discretization of the state process (x_i) , which will be seen to imply that the corresponding filtering process takes values in a finite-dimensional space. For this purpose, we start by partitioning the state space E as follows: For each positive integer m , we choose disjoint Borel sets B_k^m , $k = 1, 2, \dots, k_m$, such that

- (i) $\bigcup_{k=1}^{k_m} B_k^m = E$;
- (ii) B_k^m have nonempty interiors, the closures \bar{B}_k^m of B_k^m for $k < k_m$ are compact;
- (D_m) (iii) $\sup_{(k=1,2,\dots,k_m-1)} \text{diam}(B_k^m) \rightarrow 0$ as $m \rightarrow \infty$ where $\text{diam}(B)$ stands for the diameter of the set B ;
- (iv) $B_{k_m}^m \supset B_{k_m+1}^{m+1}$, $\bigcap_{m=1}^{\infty} B_{k_m}^m = \emptyset$;
- (v) for $k = 1, 2, \dots, k_m$, there are indices $r_1, \dots, r_{i(k)}$ such that

$$B_k^m = \bigcup_{p=1}^{i(k)} B_{r_p}^{m+1}.$$

Moreover, we assume that

(A5) Each $h^j(x)$, $j = 1, 2, \dots, d$, as well as the function $c(x, v)$ in the objective function (1.4) can be arbitrarily closely approximated in a uniform way by step functions $h_m^j(x)$ and $c_m(x, v)$ given, respectively, by

$$(3.1) \quad h_m^j(x) = \sum_{k=1}^{k_m} \chi_{B_k^m}(x) h_m^{k,j}$$

and

$$(3.2) \quad c_m(x, v) = \sum_{k=1}^{k_m} \chi_{B_k^m}(x) c_k^m(v) \quad \text{with } c_k^m \in C(U).$$

Next, for a given m , we choose a set of selectors $\{z_k^m, k = 1, 2, \dots, k_m\}$ of the partition (D_m) with the following properties:

$$(3.3) \quad \begin{aligned} z_k^m \in \text{int}(B_k^m), \quad \{z_k^m, k = 1, 2, \dots, k_m\} \subset \{z_k^{m+1}, k = 1, 2, \dots, k_{m+1}\}, \\ z_{k_m}^m \rightarrow \infty \quad \text{for } m \rightarrow \infty \end{aligned}$$

and consider the simplex S^m in \mathcal{R}^{k_m} given by

$$S^m = \left\{ (s_1, \dots, s_{k_m}), 0 \leq s_k \leq 1, k = 1, 2, \dots, k_m, \sum_{k=1}^{k_m} s_k = 1 \right\}.$$

By analogy to $\mathcal{A}(L, n)$, we let a corresponding class $\mathcal{A}_m(L, n)$ of controls $u: S^m \rightarrow U$ be defined by

$$(3.4) \quad \begin{aligned} \mathcal{A}_m(L, n) := \left\{ u \in C(S^m, U) \mid u(s) = \bar{u} \left(\sum_{k=1}^{k_m} \phi_1(z_k^m) s_k, \dots, \sum_{k=1}^{k_m} \phi_n(z_k^m) s_k \right) \right. \\ \left. \cdot r \left(\sum_{k=1}^{k_m} \psi_n(z_k^m) s_k \right) \right\}, \end{aligned}$$

where \bar{u} , ϕ_i , r , and ψ_n are as in Definition 2.1. Similarly, we have

$$(3.5) \quad \begin{aligned} \mathcal{A}_m^r(L, n) := \left\{ u \in C(S^m, U) \mid u(s) = \bar{u} \left(\sum_{k=1}^{k_m} \phi_1(z_k^m) s_k, \dots, \sum_{k=1}^{k_m} \phi_n(z_k^m) s_k \right) \right. \\ \left. \cdot r \left(\sum_{k=1}^{k_m} (\psi \wedge \psi_n)(z_k^m) s_k \right) \right\}, \end{aligned}$$

where \bar{u} , ϕ_i , r , ψ , and ψ_n are as in Definition 2.2.

Since we will have to pass from controls in one class to controls in another class, it is useful to consider the following mappings:

$$(3.6) \quad \mathcal{L}_m: \mathcal{P}(E) \ni \nu \rightarrow \sum_{k=1}^{k_m} \nu(B_k^m) \delta_{z_k^m} \in \mathcal{P}(E),$$

$$(3.7) \quad \bar{\mathcal{L}}_m: C(\mathcal{P}(E), U) \ni u \rightarrow \bar{\mathcal{L}}_m u \quad \text{with } \bar{\mathcal{L}}_m u(\nu) = u(\mathcal{L}_m \nu),$$

$$(3.8) \quad \tilde{\mathcal{L}}_m: \mathcal{A} \ni u \rightarrow \tilde{\mathcal{L}}_m u \in C(S^m, U) \quad \text{with } \tilde{\mathcal{L}}_m u(s) = u \left(\sum_{k=1}^{k_m} s_k \delta_{z_k^m} \right).$$

It follows immediately that

$$(3.9) \quad \bar{\mathcal{L}}_m u(\nu) = u \left(\sum_{k=1}^{k_m} \nu(B_k^m) \delta_{z_k^m} \right) = \tilde{\mathcal{L}}_m u(\nu(B_1^m), \dots, \nu(B_{k_m}^m))$$

and, consequently,

$$(3.10) \quad \tilde{\mathcal{L}}_m \mathcal{A}(L, n) = \mathcal{A}_m(L, n), \quad \tilde{\mathcal{L}}_m \mathcal{A}^r(l, n) = \mathcal{A}_m^r(L, n).$$

LEMMA 3.1. *We have*

$$(3.11) \quad \mathcal{L}_m \nu \Rightarrow \nu \text{ as } m \rightarrow \infty, \text{ uniformly on compact subsets of } \mathcal{P}(E);$$

$$(3.12) \quad \text{For } u \in \mathcal{A}, \bar{\mathcal{L}}_m u(\nu) \rightarrow u(\nu) \text{ as } m \rightarrow \infty, \text{ uniformly on compact subsets of } \mathcal{P}(E);$$

$$(3.13) \quad \text{If } \mathcal{B}(\mathcal{P}(E), U) \ni u_m \rightarrow u \in C(\mathcal{P}(E), U) \text{ uniformly on } \mathcal{P}(E), \text{ then } \bar{\mathcal{L}}_m u_m(\nu) \rightarrow u(\nu), \text{ uniformly on compact subsets of } \mathcal{P}(E), \text{ as } m \rightarrow \infty.$$

Proof. If $\Gamma \subset \mathcal{P}(E)$ is compact, then for any $\varepsilon > 0$ there is a compact set K and integer $m_0 > 0$ such that $K \subset \bigcup_{k=1}^p B_k^{m_0}$, $p < k_{m_0}$, and $\nu(K) > 1 - \varepsilon$ for $\nu \in \Gamma$. Then, for $\phi \in C(\mathcal{P}(E))$, using (D_m) (iii),

$$\sup_{\nu \in \Gamma} |\nu(\phi) - \mathcal{L}_m \nu(\phi)| \leq 2\varepsilon \|\phi\| + \sup_{\nu \in \Gamma} \sum_{k=1}^{k_m} \int_{K \cap B_k^m} |\phi(x) - \phi(z_k^m)| \nu(dx) \\ \rightarrow 2\varepsilon \|\phi\|, \quad \text{as } m \rightarrow \infty,$$

and, since $\varepsilon > 0$ can be chosen arbitrarily small, we obtain (3.11).

Convergence (3.12) is an almost immediate implication of (3.11) and Lemma A.1.1.

For the proof of (3.13) let us note that, denoting by ρ_U a metric compatible with the topology of U , we have

$$\rho_U(\bar{\mathcal{L}}_m u_m(\nu), u(\nu)) \leq \rho_U(\bar{\mathcal{L}}_m u_m(\nu), \bar{\mathcal{L}}_m u(\nu)) + \rho_U(\bar{\mathcal{L}}_m u(\nu), u(\nu)).$$

By (3.12) and the definition of $\bar{\mathcal{L}}_m$ we then obtain $\bar{\mathcal{L}}_m u_m(\nu) \rightarrow u(\nu)$, uniformly on compact sets of $\mathcal{P}(E)$. \square

Now let $E_m = \{1, 2, \dots, k_m\}$, which, through the correspondence $k \leftrightarrow z_k^m$, we can identify with the set $\{z_1^m, \dots, z_{k_m}^m\}$ and consider the following two approximated transition kernels:

$$(3.14) \quad \bar{P}_m^{u(\nu)}(x, \cdot) = \sum_{k=1}^{k_m} \chi_{B_k^m}(x) P^{\bar{\mathcal{L}}_m u(\nu)}(z_k^m, \cdot),$$

defined on the original state space E , and

$$(3.15) \quad P_m^v(k, p) = P^v(z_k^m, B_p^m), \quad k, p \in E_m$$

for an embedded Markov chain on E_m .

We also henceforth assume that

(A6) The partition (D_m) as well as the choice of selectors $\{z_k^m, k = 1, 2, \dots, k_m\}$ is such that, for each $k, p = 1, 2, \dots, k_m$, $v \in U$,

$$P^v(z_k^m, \partial B_p^m) = 0.$$

Note that under (A2) and (A6) the mapping $U \ni v \rightarrow P^v(z_k^m, B_p^m)$ is continuous.

Consider then the two partially observed control systems, below:

(CS₁) The unobserved state process (\bar{x}_i^m) evolves on E according to the transition operator $\bar{P}_m^u(x, \cdot)$ with initial law $\mu \in \mathcal{P}(E)$; the observations (\bar{y}_i) are of the form

$$(3.16) \quad \bar{y}_i = h_m(\bar{x}_i^m) + \bar{w}_i, \quad i = 1, 2, \dots,$$

where \bar{w}_i is a sequence of independently and identically distributed d -dimensional standard Gaussian vectors, independent of \bar{x}_k^m , $k \leq i$. We denote by $(\bar{\pi}_i^{m, \mu, u}) \in \mathcal{P}(E)$ the corresponding controlled filtering process.

(CS₂) The unobserved state process (x_i^m) evolves on E_m according to the transition probability matrix $P_m^u(k, p)$ with initial law $\bar{\mu}$ given by

$$(3.17) \quad \bar{\mu} = (\mu(B_1^m), \dots, \mu(B_{k_m}^m)) \in S^m.$$

The observations (\bar{y}_i) satisfy

$$(3.18) \quad \bar{y}_i = \sum_{k=1}^{k_m} h_m(z_k^m) \delta_{\{k\}}(x_i^m) + \bar{w}_i,$$

where, again, (\bar{w}_i) are independently and identically distributed standard Gaussian vectors independent of (x_k^m) , $k \leq i$. We denote by $(\pi_i^{m,\bar{\mu},u})$ the corresponding controlled filtering process.

By analogy to (1.32), define now, for $u \in \mathcal{A}$, $\nu \in \mathcal{P}(E)$, $y \in \mathcal{R}^d$,

$$(3.19) \quad \begin{aligned} \bar{M}_m^u(y, \nu)(\phi) &:= \int_E \int_E \exp \left[(y, h_m(z)) - \frac{1}{2} (h_m(z), h_m(z)) \right] \phi(z) \bar{P}_m^{u(\nu)}(x, dz) \nu(dx) \\ &\cdot \left(\int_E \int_E \exp \left[(y, h_m(z)) - \frac{1}{2} (h_m(z), h_m(z)) \right] \bar{P}_m^{u(\nu)}(x, dz) \nu(dx) \right)^{-1} \end{aligned}$$

and, for $u \in \mathcal{A}_m(L, n)$ or $\mathcal{A}_m^r(L, n)$, $y \in \mathcal{R}^d$, $s = (s_1, \dots, s_{k_m}) \in S^m$, $k \in E_m$,

$$(3.20) \quad \begin{aligned} M_m^u(y, s_1, \dots, s_{k_m})(k) &:= \exp \left[(y, h_m^k) - \frac{1}{2} (h_m^k, h_m^k) \right] \sum_{q=1}^{k_m} P_m^{u(s)}(q, k) s_q \\ &\cdot \left(\sum_{i=1}^{k_m} \exp \left[(y, h_m^i) - \frac{1}{2} (h_m^i, h_m^i) \right] \sum_{q=1}^{k_m} P_m^{u(s)}(q, i) s_q \right)^{-1}. \end{aligned}$$

LEMMA 3.2. For $u \in \mathcal{A}$, $i = 1, 2, \dots$, $\phi \in \mathcal{B}(E)$, we have

$$(3.21) \quad \bar{\pi}_{i+1}^{m,\mu,u}(\phi) = \bar{M}_m^u(\bar{y}_{i+1}, \bar{\pi}_i^{m,\mu,u})(\phi)$$

and $(\bar{\pi}_i^{m,\mu,u})$ is a Markov process with transition operator $(\Lambda \in \mathcal{B}(\mathcal{P}(E)))$

$$(3.22) \quad \begin{aligned} \bar{\Pi}_m^u(\nu, \Lambda) &= (2\pi)^{-(d/2)} \int_E \int_{\mathcal{R}^d} \int_E \exp \left[-\frac{1}{2} (y - h_m(\eta), y - h_m(\eta)) \right] \bar{P}_m^{u(\nu)}(\zeta, d\eta) \\ &\cdot 1_\Lambda(\bar{M}_m^u(y, \nu)) dy \nu(d\zeta). \end{aligned}$$

Furthermore, for $u \in \mathcal{A}_m(L, n)$ or $\mathcal{A}_m^r(L, n)$, $s \in S^m$, $i = 1, 2, \dots$,

$$(3.23) \quad \pi_{i+1}^{m,s,u}(k) = M_m^u(\bar{y}_{i+1}, \pi_i^{m,s,u})(k), \quad \pi_0^{m,s,u}(k) = s_k, \quad (k \in E_m)$$

and $(\pi_i^{m,s,u})$ is a Feller–Markov process with transition operator $(\Lambda \in \mathcal{B}(S))$

$$(3.24) \quad \begin{aligned} \Pi_m^u(s, \Lambda) &= (2\pi)^{-(d/2)} \sum_{k=1}^{k_m} \int_{\mathcal{R}^d} \left(\sum_{q=1}^{k_m} \exp \left[-\frac{1}{2} (y - h_m^q, y - h_m^q) \right] P_m^{u(s)}(k, q) \right) \\ &\cdot 1_\Lambda(M_m^{u(s)}(y, s)) dy s_k. \end{aligned}$$

Proof. The proof can be obtained by an adaptation of Lemma 1.1. and Proposition 1.2, using also the continuity assumption (A6) for the proof of the Feller property of $\Pi_m^u(s, F)$. \square

The actual process of interest is $(\pi_i^{m,\bar{\mu},u})$, which evolves on a finite-dimensional space; $(\bar{\pi}_i^{m,\mu,u})$ is an intermediate auxiliary process needed in the convergence proofs. To be able to state our main approximation result (namely, Theorem 3.1, below), we need some preliminary results on invariant measures for the processes $(\pi_i^{m,\bar{\mu},u})$ and $(\bar{\pi}_i^{m,\mu,u})$ that will allow us to formulate the approximating partially observable control problem on a finite-dimensional state space.

It can be easily seen that, for $u \in \mathcal{A}_m(L, n)$, the process $(\pi_i^{m, \bar{\mu}, u})$ is the Feller on the compact state space S^m ; therefore there exists an invariant measure Φ_m^u , and we have the following proposition, whose proof is given in Appendix A.2.

PROPOSITION 3.1. *Given $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$, if $\Phi_m^{\tilde{\mathcal{L}}_m u}$ is an invariant measure of $(\pi_i^{m, \bar{\mu}, \tilde{\mathcal{L}}_m u})$, then*

$$(3.25) \quad \bar{\Phi}_m^u(A) \stackrel{\text{def}}{=} \int_{S^m} \bar{\Pi}_m^u \left(\sum_{k=1}^{k_m} s_k \delta_{z_k^m}, A \right) \Phi_m^{\tilde{\mathcal{L}}_m u}(ds)$$

for $A \in \mathcal{B}(\mathcal{P}(E))$, is an invariant measure of $(\bar{\pi}_i^{m, \mu, u})$.

We now define the approximating control problem as the partially observable problem (CS₂) with cost function given, for $u \in \mathcal{A}_m(L, n)$ or $\mathcal{A}^r_m(L, n)$, by

$$(3.26) \quad \begin{aligned} J_\mu^m(u) &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_{\bar{\mu}}\{c_m(x_i^m, u(\pi_i^{m, \bar{\mu}, u}))\} \\ &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_{\bar{\mu}} \left(\sum_{k=1}^{k_m} c_k^m(u(\pi_i^{m, \bar{\mu}, u})) \pi_i^{m, \bar{\mu}, u}(k) \right), \end{aligned}$$

where, by the correspondence $k \leftrightarrow z_k^m$, we have identified $\sum_{k=1}^{k_m} z_k^m \delta_{\{k\}}(x_i^m)$ with x_i^m .

The following proposition can now be stated, whose proof is also in Appendix A.2.

PROPOSITION 3.2. *Assume (A1)–(A6) and either (B1) or (C1). Given an integer $n > 0$, let $m_0(n)$ be such that for $m > m_0(n)$ we have, in the case of assumption (B1), that $\rho(z_{k_m}^m, \bar{x}) > n$, and, more precisely, for (1.8) (respectively (1.9)),*

$$(3.27) \quad h_m^{k,j} < h_m^{k_m,j} \quad (h_m^{k,j} > h_m^{k_m,j}) \quad \text{with } k = 1, 2, \dots, k_m - 1,$$

while, in the case of (C1), there is a $q \in \{1, \dots, k_m\}$ such that $\hat{x} \in B_q^m$, $B_q^m \subset \{x: \rho(x, \hat{x}) < 1/n\}$, and, more precisely, for (1.17) (respectively, (1.18)),

$$(3.28) \quad h_m^{k,j} < h_m^{q,j} \quad (h_m^{k,j} > h_m^{q,j}) \quad \text{with } k \neq q.$$

Then, for any $u \in \mathcal{A}(L, n)$ or $u \in \mathcal{A}^r(L, n)$, according to whether (B1) or (C1) is satisfied, and $m \geq m_0(n)$ there exists a unique invariant measure Φ_m^u of $(\pi_i^{m, \bar{\mu}, u})$. Moreover, for any $u \in \mathcal{A}(L, n)$ ($\mathcal{A}^r(L, n)$), and initial law $\bar{\mu} \in S$,

$$(3.29) \quad \begin{aligned} J_\mu^m(\tilde{\mathcal{L}}_m u) &= \int_{S^m} \sum_{k=1}^{k_m} c_k^m(\tilde{\mathcal{L}}_m u(s)) s_k \Phi_m^{\tilde{\mathcal{L}}_m u}(ds) \\ &= \int_{\mathcal{P}(E)} \int_E c_m(x, \tilde{\mathcal{L}}_m u(\nu)) \nu(dx) \bar{\Phi}_m^u(d\nu) \end{aligned}$$

with $\bar{\Phi}_m^u$ defined by (3.25).

We are now ready to prove the main state-approximation result.

THEOREM 3.1. *Assume (A1)–(A6) together with (B1), (B4), or (C1), (C2), or (C1) and the compactness of E . Then, under (B1), (B4),*

$$(3.30) \quad \lim_{m \rightarrow \infty} \inf_{u \in \mathcal{A}_m(L, n)} J_\mu^m(u) = \inf_{u \in \mathcal{A}(L, n)} J_\mu(u),$$

while under (C1), (C2), or (C1) and the compactness of E ,

$$(3.31) \quad \lim_{m \rightarrow \infty} \inf_{u \in \mathcal{A}_m^r(L, n)} J_\mu^m(u) = \inf_{u \in \mathcal{A}^r(L, n)} J_\mu(u)$$

for each $\mu \in \mathcal{P}(E)$. Moreover, if for $s \in S^m$, and $\varepsilon > 0$

$$(3.32) \quad \begin{aligned} \tilde{u}_m(s_1, \dots, s_{k_m}) &= \bar{u}_m \left(\sum_{k=1}^{k_m} \phi_1(z_k^m) s_k, \dots, \sum_{k=1}^{k_m} \phi_n(z_k^m) s_k \right) \\ &\quad \cdot r \left(\sum_{k=1}^{k_m} \psi_n(z_k^m) s_k \right), \end{aligned}$$

$$(3.33) \quad \begin{aligned} \tilde{u}_m(s_1, \dots, s_{k_m}) &= \bar{u}_m \left(\sum_{k=1}^{k_m} \phi_1(z_k^m) s_k, \dots, \sum_{k=1}^{k_m} \phi_n(z_k^m) s_k \right) \\ &\quad \cdot r \left(\sum_{k=1}^{k_m} (\psi^m \wedge \psi_n)(z_k^m) s_k \right) \end{aligned}$$

are for $m=1, 2, \dots$ sequences of ε -optimal control functions for $J_{\bar{\mu}}^m(u)$ over $\mathcal{A}_m(L, n)(\mathcal{A}'_m(L, n)$, respectively), then any uniform limit of

$$(3.34) \quad \hat{u}_m(v) = \bar{u}_m(v(\phi_1), \dots, v(\phi_n)) r(v(\psi_n)),$$

$$(3.35) \quad \hat{u}_m(v) = \bar{u}_m(v(\phi_1), \dots, v(\phi_n)) r(v(\psi^m \wedge \psi_n))$$

forms an ε -optimal control function for $J_{\bar{\mu}}(u)$ over $\mathcal{A}(L, n)(\mathcal{A}'(L, n)$, respectively).

Remark 3.1. In Theorem 3.1, we make assumptions (A1)–(A6) and one among the three alternative additional groups (B1), (B4), or (C1), (C2), or (C1) and the compactness of E . This setting will remain the same for the rest of the paper. We want to point out, however, that the results to follow can also be obtained under assumptions (B1), (B3) by imposing additional tightness conditions; this latter approach is taken in [13].

Proof of Theorem 3.1. We divide the proof into two steps.

Step 1. For a given strategy $u \in \mathcal{A}(L, n)$ or $\mathcal{A}'(L, n)$, we have

$$(3.36) \quad \lim_{m \rightarrow \infty} J_{\bar{\mu}}^m(\tilde{\mathcal{L}}_m u) = J_{\bar{\mu}}(u).$$

For this purpose, letting

$$(3.37) \quad \tilde{P}_m^\nu(x, \cdot) = \sum_{k=1}^{k_m} \chi_{B_k^m}(x) P^\nu(z_k^m, \cdot),$$

note that all assumptions of Theorem 2.1 or Theorem 2.2 are satisfied with $u_m = \tilde{\mathcal{L}}_m u$, $\Phi_m^{u_m} = \bar{\Phi}_m^u$, $\pi_i^{m, \mu, u_m} = \bar{\pi}_i^{m, \mu, u}$, $y_i = \bar{y}_i$, $P_m^{u_m} = \bar{P}_m^u$, $P_m^v = \tilde{P}_m^v$. In fact, under our assumptions, by the second part of Theorem 2.3, we have that, for each $u \in \mathcal{A}(L, n)$ or $\mathcal{A}'(L, n)$, there exists a unique invariant measure Φ^u of the filtering process $(\pi_i^{\mu, u})$. Furthermore, (2.1) follows from (3.38) below, (2.2) is a consequence of Lemma 3.1, (2.3) is guaranteed by (A.5), and (B4) holds, uniformly in m , for each $P_m^v(x, \cdot)$ by its definition. We next show that under our assumptions we have

$$(3.38) \quad \begin{aligned} &\text{if } U \ni v_m \rightarrow v, \text{ then} \\ &\tilde{P}_m^{v_m} \Rightarrow P^v(x, \cdot) \text{ uniformly in } x \text{ from compact subsets of } E. \end{aligned}$$

In fact, assume that $K \subset E$ is a compact set. By the construction of the partition (D_m) , for each m and $x \in K$, there exists an index $m(x)$ such that $x \in B_{m(x)}^m$ and, letting $m \rightarrow \infty$, $\sup_{x \in K} \rho(z_{m(x)}^m, x) \rightarrow 0$. Then, for $f \in C(E)$ we have from (A1) and (A2)

$$(3.39) \quad \begin{aligned} |\tilde{P}_m^{v_m} f(x) - P^v f(x)| &\leq |P_m^{v_m} f(z_{m(x)}^m) - P^v f(z_{m(x)}^m)| \\ &\quad + |P^v f(z_{m(x)}^m) - P^v f(x)| \rightarrow 0 \quad \text{for } m \rightarrow \infty, \end{aligned}$$

from which (3.38) follows. Applying Theorem 2.1 or Theorem 2.2, we then have

$$(3.40) \quad \bar{\Phi}_m^u \Rightarrow \Phi^u \quad \text{for } m \rightarrow \infty,$$

and for any $\varepsilon > 0$ there is a compact set $\Gamma(\varepsilon) \subset \mathcal{P}(E)$ such that for $m = 1, 2, \dots$

$$(3.41) \quad \bar{\Phi}_m^u(\Gamma(\varepsilon)) \geq 1 - \varepsilon, \quad \Phi^u(\Gamma(\varepsilon)) \geq 1 - \varepsilon.$$

Since $\Gamma(\varepsilon)$ is tight, there is a compact set $K(\varepsilon) \subset E$ such that

$$(3.42) \quad \nu(K(\varepsilon)) \geq 1 - \varepsilon \quad \text{for } \nu \in \Gamma(\varepsilon).$$

By the continuity of $c(x, v)$, given $\varepsilon > 0$, there is $\delta > 0$ such that

$$(3.43) \quad \rho_U(v, v') < \delta \quad \text{implies} \quad \sup_{x \in K(\varepsilon)} |c(x, v) - c(x, v')| < \varepsilon.$$

Finally, by Lemma 3.1, for $m > \bar{m}$

$$(3.44) \quad \sup_{\nu \in \Gamma(\varepsilon)} \rho_U(\bar{\mathcal{L}}_m u(\nu), u(\nu)) < \delta.$$

Now, applying (3.29), we have

$$(3.45) \quad \begin{aligned} |J_\mu(u) - J_\mu^m(\bar{\mathcal{L}}_m u)| &= \left| \int_{\mathcal{P}(E)} \int_E c(x, u(\nu)) \nu(dx) \Phi^u(d\nu) \right. \\ &\quad \left. - \int_{\mathcal{P}(E)} \int_E c_m(x, \bar{\mathcal{L}}_m(u(\nu))) \nu(dx) \bar{\Phi}_m^u(d\nu) \right| \\ &\leq \left| \int_{\mathcal{P}(E)} \int_E c(x, u(\nu)) \nu(dx) (\Phi^u(d\nu) - \bar{\Phi}_m^u(d\nu)) \right| \\ &\quad + \left| \int_{\mathcal{P}(E)} \int_E (c(x, u(\nu)) - c(x, \bar{\mathcal{L}}_m u(\nu))) \nu(dx) \bar{\Phi}_m^u(d\nu) \right| \\ &\quad + \left| \int_{\mathcal{P}(E)} \int_E (c(x, \bar{\mathcal{L}}_m u(\nu)) - c_m(x, \bar{\mathcal{L}}_m u(\nu))) \nu(dx) \bar{\Phi}_m^u(d\nu) \right| \\ &= \text{I}_m + \text{II}_m + \text{III}_m. \end{aligned}$$

Clearly, (3.40) implies that $\text{I}_m \rightarrow 0$, and from (A5) it follows that $\text{III}_m \rightarrow 0$. By (3.41)–(3.44) for $m > \bar{m}$,

$$(3.46) \quad \begin{aligned} \text{II}_m &\leq \int_{\Gamma(\varepsilon)} \int_{K(\varepsilon)} |c(x, u(\nu)) - c(x, \bar{\mathcal{L}}_m u(\nu))| \nu(dx) \bar{\Phi}_m^u(d\nu) + 4\|c\|\varepsilon \\ &\leq \varepsilon(1 + 4\|c\|). \end{aligned}$$

Thus (3.36) holds.

Step 2. From Step 1, we have

$$(3.47) \quad \limsup_{m \rightarrow \infty} \inf_{u \in \mathcal{A}'_m(L, n)} J_\mu^m(u) \leq \inf_{u \in \mathcal{A}'(L, n)} J_\mu(u)$$

or

$$(3.48) \quad \limsup_{m \rightarrow \infty} \inf_{u \in \mathcal{A}'_m(L, n)} J_\mu^m(u) \leq \inf_{u \in \mathcal{A}'(L, n)} J_\mu(u).$$

The inverse inequalities follow from the second assertion of the theorem, an assertion that we will prove next. Let \tilde{u}_m , given by (3.32) (respectively, (3.33)), be such that for $m = 1, 2, \dots$, they form a sequence of ε -optimal control functions for $J_\mu^m(u)$ over $\mathcal{A}_m(L, n)$ ($\mathcal{A}'_m(L, n)$, respectively). Since \tilde{u}_m are bounded and Lipschitz with Lipschitz

constant L , by Ascoli's theorem (see, e.g., [11, Thm. 9.33]), there is $\bar{u} \in C(R^n, U)$ and a subsequence $m_k \rightarrow \infty$ such that $\bar{u}_{m_k} \rightarrow \bar{u}$ uniformly on $[-\|\phi_1\|, \|\phi_1\|] \times \cdots \times [-\|\phi_n\|, \|\phi_n\|]$. Also, since ψ^m are Lipschitz with Lipschitz constant n and values from $[0, 1]$, we may assume that m_k is chosen in such a way that $\psi^{m_k}(x) \rightarrow \psi(x)$ uniformly in K_n . Let us now define

$$(3.49) \quad \hat{u}(\nu) = \bar{u}(\nu(\phi_1), \dots, \nu(\phi_n))r(\nu(\psi_n)) \in \mathcal{A}(L, n)$$

or

$$(3.50) \quad \hat{u}(\nu) = \bar{u}(\nu(\phi_1), \dots, \nu(\phi_n))r(\nu(\psi \wedge \psi_n)) \in \mathcal{A}^r(L, n).$$

Then, both in the case $\tilde{u}_{m_k} \in \mathcal{A}(L, n)$ and $\tilde{u}_{m_k} \in \mathcal{A}^r(L, n)$, we have

$$(3.51) \quad \hat{u}_{m_k}(\nu) \rightarrow \hat{u}(\nu) \quad \text{uniformly in } \nu \in \mathcal{P}(E)$$

with $\hat{u}(\nu)$ given by (3.49) or (3.50), accordingly. Hence, by (3.13), $\bar{\mathcal{L}}_{m_k} \hat{u}_{m_k}(\nu) \rightarrow \hat{u}(\nu)$ uniformly on compact subsets of $\mathcal{P}(E)$. Letting the subsequence m_k also be denoted by m for simplicity of notation, we again apply Theorem 2.2, this time with $u = \hat{u}$, $u_m = \bar{\mathcal{L}}_m \hat{u}_m$, $\Phi_m^u = \bar{\Phi}_m^{\hat{u}_m}$, $\pi_i^{m, \mu, u_m} = \bar{\pi}_i^{m, \mu, \hat{u}_m}$, $P_m^u = \bar{P}_m^u$, $P_m^v = \bar{P}_m^v$. Therefore

$$(3.52) \quad \bar{\Phi}_m^{\hat{u}_m} \Rightarrow \bar{\Phi}^{\hat{u}},$$

and, by analogy to (3.41)–(3.46), we obtain

$$(3.53) \quad \lim_{m \rightarrow \infty} J_{\bar{\mu}}^m(\bar{\mathcal{L}}_m \hat{u}_m) = J_{\mu}(\hat{u}).$$

Since $\tilde{u}_m = \bar{\mathcal{L}}_m \hat{u}_m$, as can easily be checked, and \tilde{u}_m is ε -optimal for $J_{\bar{\mu}}^m$ over $\mathcal{A}_m(L, n)$ ($\mathcal{A}_m^r(L, n)$, respectively) from (3.47) and (3.48), we have

$$(3.54) \quad J_{\mu}(\hat{u}) \leq \limsup_{m \rightarrow \infty} \inf_{u \in \mathcal{A}_m(L, n)} J_{\bar{\mu}}^m(u) + \varepsilon \leq \inf_{u \in \mathcal{A}(L, n)} J_{\mu}(u) + \varepsilon$$

and

$$(3.55) \quad J_{\mu}(\hat{u}) \leq \limsup_{m \rightarrow \infty} \inf_{u \in \mathcal{A}_m^r(L, n)} J_{\bar{\mu}}^m(u) + \varepsilon \leq \inf_{u \in \mathcal{A}^r(L, n)} J_{\mu}(u) + \varepsilon,$$

respectively. This means, however, that \hat{u} given by (3.49) (respectively, (3.50)) is ε -optimal for J_{μ} over $\mathcal{A}(L, n)$, ($\mathcal{A}^r(L, n)$, respectively), which is the claim of the second part of the theorem. In addition, since in (3.54), (3.55) ε can be chosen arbitrarily small, taking also (3.47) and (3.48) into account, we obtain (3.30) ((3.31), respectively). The proof of Theorem 3.1 is thus completed. \square

COROLLARY 3.1. *Under the assumptions of Theorem 3.1, if \tilde{u}_m given by (3.32) (respectively, (3.33)) are ε -optimal for $J_{\bar{\mu}}^m(u)$ over $\mathcal{A}_m(L, n)$ (respectively, $\mathcal{A}_m^r(L, n)$), then \hat{u}_m defined in (3.34) (respectively, (3.35)) are, for m sufficiently large, 2ε -optimal for $J_{\mu}(u)$ over $\mathcal{A}(L, n)$ (respectively, $\mathcal{A}^r(L, n)$).*

Proof. It is enough to show that

$$(3.56) \quad \limsup_{m \rightarrow \infty} J_{\mu}(\hat{u}_m) \leq \inf_{u \in \mathcal{A}(L, n)} J_{\mu}(u) + \varepsilon,$$

which follows from the fact that, for any subsequence \hat{u}_{m_k} converging uniformly to \hat{u} , again from Theorem 2.1 we have

$$(3.57) \quad J_{\mu}(\hat{u}_{m_k}) \rightarrow J_{\mu}(\hat{u}) \leq \inf_{u \in \mathcal{A}(L, n)} J_{\mu}(u) + \varepsilon. \quad \square$$

3.2. Finite set of control functions. By Theorem 3.1 and Corollary 3.1, we know that we can obtain a nearly optimal control function for our original problem in the classes $\mathcal{A}(L, n)$, $\mathcal{A}^r(L, n)$, provided that we are able to obtain, for sufficiently large m , an optimal or at least nearly optimal control function for $J_{\bar{\mu}}^m(u)$ in the classes

$\mathcal{A}_m(L, n)$ or $\mathcal{A}'_m(L, n)$, respectively. By Theorem 2.3, such control function is then nearly optimal also for $J_\mu(u)$ in the original classes \mathcal{A} or $\bar{\mathcal{A}}$ and \mathcal{A}' , respectively, if L and n are sufficiently large. The construction of an optimal control function in the classes $\mathcal{A}_m(L, n)$, $\mathcal{A}'_m(L, n)$ is still a formidable problem; on the other hand, we need only a nearly optimal control function, and this allows us to proceed with further approximations. For this purpose, note first that the sets of Lipschitz-continuous functions $C_L(H_n, U)$ from $H_n := [-\|\phi_1\|, \|\phi_1\|] \times \cdots \times [-\|\phi_n\|, \|\phi_n\|] \subset \mathcal{R}^n$ into the compact set $U \subset \mathcal{R}^k$, and $C_n(K_n, [0, 1])$ from $K_n = \{x: \rho(x, \bar{x}) \leq n\}$ into $[0, 1]$, are compact, so that for any $\delta > 0$ there exist finite δ -nets $C_L(\delta) \subset C_L(H_n, U)$ and $\tilde{C}_n(\delta) \subset C_n(K_n, [0, 1])$. Since the functions $\bar{u}: \mathcal{R}^n \rightarrow U$, used in the definitions of $\mathcal{A}_m(L, n)$ and $\mathcal{A}'_m(L, n)$, are Lipschitz with constant L and, for given ϕ_1, \dots, ϕ_n , they are actually defined on the compact set H_n , consider then the classes $\mathcal{A}_m^{\delta}(L, n) \subset \mathcal{A}_m(L, n)$, and $\mathcal{A}_m^{r, \delta}(L, n) \subset \mathcal{A}'_m(L, n)$ given by

$$(3.58) \quad \mathcal{A}_m^{\delta}(L, n) = \left\{ u \in C(S^m, U): \exists \bar{u} \in C_L(\delta) \text{ for which} \right. \\ \left. u(s) = \bar{u} \left(\sum_{k=1}^{k_m} \phi_1(z_k^m) s_k, \dots, \sum_{k=1}^{k_m} \phi_n(z_k^m) s_k \right) \right. \\ \left. \cdot r \left(\sum_{k=1}^{k_m} \psi_n(z_k^m) s_k \right) \right\},$$

$$(3.59) \quad \mathcal{A}_m^{r, \delta}(L, n) = \left\{ u \in C(S^m, U): \exists \bar{u} \in C_L(\delta), \psi \in \tilde{C}_n(\delta) \text{ for which} \right. \\ \left. u(s) = \bar{u} \left(\sum_{k=1}^{k_m} \phi_1(z_k^m) s_k, \dots, \sum_{k=1}^{k_m} \phi_n(z_k^m) s_k \right) \right. \\ \left. \cdot r \left(\sum_{k=1}^{k_m} (\psi \wedge \psi_n)(z_k^m) s_k \right) \right\}$$

and note that such classes contain only a finite number of elements. In principle, it is then possible to evaluate $J_\mu^m(u)$ for each $u \in \mathcal{A}_m^{\delta}(L, n)$, $\mathcal{A}_m^{r, \delta}(L, n)$, compare the corresponding values, and thus determine an optimal control function u_δ^* for $J_\mu^m(u)$ in the classes $\mathcal{A}_m^{\delta}(L, n)$ and $\mathcal{A}_m^{r, \delta}(L, n)$, respectively. To evaluate $J_\mu^m(u)$, by Propositions 3.1 and 3.2, we must be able to compute the invariant measure Φ_m^u of the process $(\pi_i^{m, \bar{\mu}, u})$ on S^m for each $u \in \mathcal{A}_m^{\delta}(L, n)$ or $\mathcal{A}_m^{r, \delta}(L, n)$. Since S^m is infinite-valued, to make such computation feasible, we must introduce further approximations, so that instead of an optimal u_δ^* we will be able to compute only an ε -optimal control function u_δ^ε for $J_\mu^m(u)$ in $\mathcal{A}_m^{\delta}(L, n)$ and $\mathcal{A}_m^{r, \delta}(L, n)$, respectively. The details of such further approximations are fully described in [13]; here we limit ourselves to prove Theorem 3.2, which shows that, for $\delta > 0$ sufficiently small, the computation of u_δ^ε in the classes $\mathcal{A}_m^{\delta}(L, n)$ and $\mathcal{A}_m^{r, \delta}(L, n)$ is sufficient for our purposes; all we need, in fact, is a nearly optimal control function for $J_\mu^m(u)$ in the classes $\mathcal{A}_m(L, n)$ and $\mathcal{A}'_m(L, n)$, respectively.

THEOREM 3.2. *Assume (A1)–(A6) and either (B1) or (C1). For fixed L and positive integer n , we then have, for every $m > m_0(n)$, where $m_0(n)$ is given in Proposition 3.2, and any $\bar{\mu} \in S^m$,*

$$(3.60) \quad \lim_{\delta \rightarrow 0} \inf_{u \in \mathcal{A}_m^{\delta}(L, n)} J_\mu^m(u) = \inf_{u \in \mathcal{A}_m(L, n)} J_\mu^m(u)$$

for the case of (B1), and

$$(3.61) \quad \lim_{\delta \rightarrow 0} \inf_{u \in \mathcal{A}_m^{r, \delta}(L, n)} J_\mu^m(u) = \inf_{u \in \mathcal{A}_m^{r, \delta}(L, n)} J_\mu^m(u)$$

for the case of (C1).

Proof. With $E = E_m$, $\mathcal{P}(E) = S^m$, and since, furthermore, letting $\delta \rightarrow 0$, each $u \in \mathcal{A}_m(L, n)$ or $\mathcal{A}_m^r(L, n)$ can be uniformly approximated by $u_\delta \in \mathcal{A}_m^\delta(L, n)$ ($\mathcal{A}_m^{r,\delta}(L, n)$, respectively), all assumptions of Theorem 2.1 are satisfied. Therefore, by Theorem 2.1 we have $J_\mu^m(u) = \lim_{\delta \rightarrow 0} J_\mu^m(u_\delta)$, and consequently (3.60) and (3.61) hold. \square

3.3. Algorithmic structure of our procedure. Summarizing the results of § 3, we have a constructive method for determining a nearly optimal control function for the original problem. This control function is given by an ε -optimal control function u_δ^ε for $J_\mu^m(u)$ over $\mathcal{A}_m^\delta(L, n)$ and $\mathcal{A}_m^{r,\delta}(L, n)$, respectively, which, by the further approximations described in [13], can indeed be computed. In fact, by Theorem 3.2, such u_δ^ε is 2ε -optimal for J_μ^m over $\mathcal{A}_m(L, n)$ (respectively, $\mathcal{A}_m^r(L, n)$) if δ is sufficiently small. Combining this finally with the results of §§ 2.2 and 3.1, we have that, provided furthermore that L , n and m are sufficiently large, this u_δ^ε induces a nearly optimal control function $\hat{u}_\delta^\varepsilon \in \mathcal{A}(L, n)$ (respectively, $\mathcal{A}^r(L, n)$) for the original problem. To obtain from this control function $\hat{u}_\delta^\varepsilon$ the actual control values, it must be evaluated for the current filter values $(\pi_i^{\mu,u})$ of the original problem. Although such filter values cannot be computed in practice, we are nevertheless able to compute (see (4.6) below) an approximating filtering process $(\pi_i^{m(\bar{\mu})})$ over the simplex S^m . In § 4 we show that, by using $\hat{u}_\delta^\varepsilon$ with the approximate filter values $\pi_i^{m(\bar{\mu})}$, we indeed obtain (for L , n , m , sufficiently large and δ sufficiently small) nearly optimal controls for the original problem.

4. Filter approximation and near optimal control values. The purpose of this section is to show that (provided that the values of the parameters L , n in § 2.2, and m in § 3.1 are sufficiently large, while that of δ in § 3.2 is sufficiently small) the control function $\hat{u}_\delta^\varepsilon \in \mathcal{A}(L, n) \subset \mathcal{A}$ (or $\mathcal{A}^r(L, n) \subset \mathcal{A}^r$), determined according to the previous § 3, yields nearly optimal control values also if evaluated in correspondence of the computable approximate filter $(\pi_i^{m(\bar{\mu})}) \in S^m$ defined in (4.6) below for a generic initial measure $\bar{\nu} \in S^m$.

To obtain the desired result, we must also introduce, besides $(\pi_i^{m(\bar{\mu})})$, other auxiliary measure-valued processes. For this purpose, recall the measure $\zeta^u(\mu)(\cdot)$ defined in (2.17) and analogously let

$$(4.1) \quad \zeta^u(\mu, \nu)(\cdot) = \int_E P^{u(\nu)}(x, \cdot) \mu(dx)$$

for any $\mu, \nu \in \mathcal{P}(E)$ and $u \in \mathcal{A}$.

Given $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$ and an arbitrary measure $\nu \in \mathcal{P}(E)$, first let $(\pi_i^{(\nu)})$ be a process in $\mathcal{P}(E)$ defined recursively by

$$(4.2) \quad \pi_0^{(\nu)}(\cdot) = \nu(\cdot), \quad \pi_{i+1}^{(\nu)}(\cdot) = R(y_{i+1}^{(\nu)}, \zeta^u(\pi_i^{(\nu)}))(\cdot),$$

where $R(\cdot, \cdot)$ is as defined in (2.15) and $(y_i^{(\nu)})$ are the observations given by (1.1) when (x_i) is governed by the transition function $P^{u(\pi_i^{(\nu)})}(\cdot, \cdot)$. Furthermore, given $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$ and $\mu, \nu \in \mathcal{P}(E)$, let the process $(\pi_i^{(\mu, \nu)})$ in $\mathcal{P}(E)$ be defined by

$$(4.3) \quad \pi_0^{(\mu, \nu)}(\cdot) = \mu(\cdot), \quad \pi_{i+1}^{(\mu, \nu)}(\cdot) = R(y_{i+1}^{(\nu)}, \zeta^u(\pi_i^{(\mu, \nu)}, \pi_i^{(\nu)}))(\cdot),$$

where $(y_i^{(\nu)})$ are as above.

Note that, if the state process (x_i) is governed by the same transition function $P^{u(\pi_i^{(\nu)})}(\cdot, \cdot)$ as above, but starts with a given initial law μ , then by (1.30)–(1.32) we obtain that

$$(4.4) \quad E_\mu\{\phi(x_i) | Y^i\} = R(y_i^{(\nu)}, \zeta^u(\pi_{i-1}^{(\mu, \nu)}, \pi_{i-1}^{(\nu)}))(\phi) = \pi_i^{(\mu, \nu)}(\phi) \quad \text{for } i = 1, 2, \dots$$

Given $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$, $\nu \in \mathcal{P}(E)$, $\bar{\nu} \in S^m$, with $\bar{\nu}(k) = \nu(B_k^m)$ for $k = 1, 2, \dots, k_m$, by analogy to (3.21), (3.23), we further introduce two approximating filtering processes $\bar{\pi}_i^{m(\nu)}$ in $\mathcal{P}(E)$ and $\pi_i^{m(\bar{\nu})}$ in S^m defined as follows:

$$(4.5) \quad \bar{\pi}_0^{m(\nu)}(\cdot) = \nu(\cdot), \quad \bar{\pi}_{i+1}^{m(\nu)}(\cdot) = \bar{M}_m^u(y_{i+1}^{(\nu)}, \bar{\pi}_i^{m(\nu)})(\cdot);$$

$$(4.6) \quad \pi_0^{m(\bar{\nu})}(k) = \bar{\nu}(k), \quad \pi_{i+1}^{m(\bar{\nu})}(k) = M_m^{\bar{\mathcal{F}}_m^u}(y_{i+1}^{(\nu)}, \pi_i^{m(\bar{\nu})})(k),$$

with $y_i^{(\nu)}$ now denoting the observations given by (1.1) when the state process (x_i) is governed by the transition function $P_m^u(\bar{\pi}_i^{m(\nu)})(\cdot, \cdot)$, which, due to (3.9) and the fact that $\bar{\pi}_i^{m(\nu)}(B_k^m) = \pi_i^{m(\bar{\nu})}(k)$ (cf. Lemma A.2.1), is equivalent to $P_m^{\bar{\mathcal{F}}_m^u}(\pi_i^{m(\bar{\nu})})(\cdot, \cdot)$. Note that the approximating filter process $(\pi_i^{m(\bar{\nu})}) \in S^m$ is explicitly computable; in fact, the observations $y_i^{(\nu)}$ correspond to our real observations since (x_i) is governed by the original transition kernel $P^v(\cdot, \cdot)$ with control values $v_i = \bar{\mathcal{L}}_m^u(\pi_i^{m(\bar{\nu})})$.

Finally, given $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$ and $\mu, \nu \in \mathcal{P}(E)$, let, by analogy to (4.3), the process $(\pi_i^{m(\mu, \nu)})$ in $\mathcal{P}(E)$ be defined by

$$(4.7) \quad \begin{aligned} \pi_0^{m(\mu, \nu)}(\cdot) &= \mu(\cdot), \\ \pi_{i+1}^{m(\mu, \nu)}(\cdot) &= R(y_{i+1}^{(\nu)}, \zeta^{\bar{\mathcal{F}}_m^u}(\pi_i^{m(\mu, \nu)}, \pi_i^{m(\bar{\nu})}))(\cdot) \\ &= R(y_{i+1}^{(\nu)}, \zeta^{\bar{\mathcal{F}}_m^u}(\pi_i^{m(\mu, \nu)}, \bar{\pi}_i^{m(\nu)}))(\cdot), \end{aligned}$$

where $(y_i^{(\nu)})$ is as defined below (4.6) and where we have implicitly extended the definition of $\zeta^u(\cdot, \cdot)$, again using the fact that $\bar{\pi}_i^{m(\nu)}(B_k^m) = \pi_i^{m(\bar{\nu})}(k)$. Analogously to (4.4), note again that, if the state process (x_i) is governed by the transition function $P_m^{\bar{\mathcal{F}}_m^u}(\pi_i^{m(\nu)})(\cdot, \cdot) = P_m^{\bar{\mathcal{F}}_m^u}(\pi_i^{m(\nu)})(\cdot, \cdot)$ and starts with the given initial law μ , then by (1.30)–(1.32) we obtain that

$$(4.8) \quad E_\mu\{\phi(x_i) | Y^i\} = \pi_i^{m(\mu, \nu)}(\phi) \quad \text{for } i = 1, 2, \dots$$

Therefore, the process $\pi_i^{m(\mu, \nu)}$ depends on the measure ν through the values of $\bar{\nu} = (\nu(B_1^m), \dots, \nu(B_{k_m}^m))$ only. With abuse of notation, we write $\pi_i^{m(\mu, \nu)} = \pi_i^{m(\mu, \bar{\nu})}$ and define the process $\pi_i^{m(\mu, s)}$ with $s \in S^m$ as equal to $\pi_i^{m(\mu, \nu)}$ with $\bar{\nu} = s$. (For the relationship between ν and $\bar{\nu}$, we recall (3.17).) We are now in a position to consider the three pairs of processes $(\pi_i^{(\mu, \nu)}, \pi_i^{(\nu)})$, $(\pi_i^{m(\mu, \nu)}, \bar{\pi}_i^{m(\nu)})$, $(\pi_i^{m(\mu, \nu)}, \pi_i^{m(\bar{\nu})})$. By considerations similar to those in the proof of Proposition 1.2, we can show that these processes are Markov with transition operators T^u , \bar{T}_m^u , T_m^u defined, respectively, as follows: For $F \in \mathcal{B}(\mathcal{P}(E) \times \mathcal{P}(E))$ the set of bounded Borel functions on $\mathcal{P}(E) \times \mathcal{P}(E)$ and $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$, we have

$$(4.9) \quad \begin{aligned} T^u F(\mu, \nu) &= E_\mu\{F(\pi_1^{(\mu, \nu)}, \pi_1^{(\nu)})\} \\ &= \int_E \int_E \left(\int_{\mathcal{R}^d} (2\pi)^{-d/2} \exp \left[-\frac{1}{2} (y - h(z), y - h(z)) \right] \right) \\ &\quad \cdot F(R(y, \zeta^u(\mu, \nu)), R(y, \zeta^u(\nu))) dy P^{u(\nu)}(x, dz) \mu(dx), \end{aligned}$$

$$(4.10) \quad \begin{aligned} \bar{T}_m^u F(\mu, \nu) &= E_\mu\{F(\pi_1^{m(\mu, \nu)}, \bar{\pi}_1^{m(\nu)})\} \\ &= \int_E \int_E \left(\int_{\mathcal{R}^d} (2\pi)^{-d/2} \exp \left[-\frac{1}{2} (y - h(z), y - h(z)) \right] \right) \\ &\quad \cdot F(R(y, \zeta^{\bar{\mathcal{F}}_m^u}(\mu, \nu)), \bar{M}_m^u(y, \nu)) dy P^{\bar{\mathcal{F}}_m^u}(\nu)(x, dz) \mu(dx), \end{aligned}$$

while for $\bar{F} \in \mathcal{B}(\mathcal{P}(E) \times S^m)$

$$\begin{aligned}
 T_m^u \bar{F}(\mu, \bar{\nu}) &= E_\mu \{ \bar{F}(\pi_1^{m(\mu, \bar{\nu})}, \pi_1^{m(\bar{\nu})}) \} \\
 (4.11) \quad &= \int_E \int_E \left(\int_{\mathcal{R}^d} (2\pi)^{-d/2} \exp \left[-\frac{1}{2} (y - h(z), y - h(z)) \right] \right. \\
 &\quad \cdot \bar{F}(R(y, \zeta^{\tilde{\mathcal{F}}_m^u}(\mu, \bar{\nu})), M_m^{\tilde{\mathcal{F}}_m^u}(y, \bar{\nu})) dy \\
 &\quad \cdot P^{\tilde{\mathcal{F}}_m^u(\bar{\nu})}(x, dz) \mu(dx).
 \end{aligned}$$

Moreover, by a suitable adaptation of the proof of Proposition 1.2, we can show that under (A1), (A2) the operator T^u is Feller. Assuming, in addition, (A6), we obtain the Feller property of T_m^u as well.

The relationship between the operators \bar{T}_m^u and T_m^u is shown in Lemmas 4.1 and 4.2. By analogy to Lemma A.2.1, we first have Lemma 4.1, whose proof is completely similar to the second part of that of Lemma A.2.1.

LEMMA 4.1. *Let $F \in \mathcal{B}(\mathcal{P}(E) \times S^m)$, then, for $\mu, \nu \in \mathcal{P}(E)$,*

$$\begin{aligned}
 (4.12) \quad &\int_{\mathcal{P}(E)} \int_{\mathcal{P}(E)} F(\mu', \nu'(B_1^m), \dots, \nu'(B_{k_m}^m)) \bar{T}_m^u(\mu, \nu, d\mu' \times d\nu') \\
 &= \int_{\mathcal{P}(E)} \int_{S^m} F(\mu', s_1, \dots, s_{k_m}) T_m^u(\mu, \bar{\nu}, d\mu' \times ds).
 \end{aligned}$$

Corresponding to Proposition 3.1 and using Lemma 4.1, we immediately have Lemma 4.2 as well.

LEMMA 4.2. *If, for $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$ the measure Ψ_m^u is an invariant measure of the operator T_m^u , then with $A \in \mathcal{B}(\mathcal{P}(E) \times \mathcal{P}(E))$*

$$(4.13) \quad \bar{\Psi}_m^u(A) \stackrel{\text{def}}{=} \int_{\mathcal{P}(E) \times S^m} \bar{T}_m^u \left(\mu', \sum_{k=1}^{k_m} s_k \delta_{z_k^m}, A \right) \Psi_m^u(d\mu', ds)$$

is an invariant measure for the transition operator \bar{T}_m^u .

PROPOSITION 4.1. *Assume (A1)–(A4). Then*

- (i) *Under (B1), (B4), for each $u \in \mathcal{A}(L, n)$, there is a unique invariant measure Ψ^u of T^u ;*
- (ii) *Under (C1), (C2), for each $u \in \mathcal{A}^r(L, n)$, there is a unique invariant measure Ψ^u of T^u ;*
- (iii) *Under (A6) and (B1), (B2), for each $u \in \mathcal{A}(L, n)$ and $m > m_0(n)$ with $m_0(n)$ defined in Proposition 3.2, there is a unique invariant measure Ψ_m^u of T_m^u ;*
- (iv) *Under (A6) and (C1), (C2), for each $u \in \mathcal{A}^r(L, n)$, and $m > m_0(n)$, there is a unique invariant measure Ψ_m^u of T_m^u ;*
- (v) *Under (A6), (C1) and the compactness of E , for each $u \in \mathcal{A}^r(L, n)$ and $m > m_0(n)$, there exists a unique invariant measure Ψ^u of T^u and Ψ_m^u of T_m^u .*

Proof. For cases (i)–(iv), we use the arguments from the proof of Theorem 2.2; namely, letting

$$\begin{aligned}
 \bar{\Gamma} &= \{(\mu, \nu) \in \mathcal{P}(E) \times \mathcal{P}(E) \mid \nu(\bar{\psi}_n) < b\}, \\
 \tilde{\Gamma} &= \left\{ (\mu, s) \in \mathcal{P}(E) \times S^m \mid \sum_{k=1}^{k_m} s_k \psi_n(z_k^m) < b \right\},
 \end{aligned}$$

define

$$\begin{aligned}
 \tau &= \inf \{i > 0 \mid (\pi_i^{(\mu, \nu)}, \pi_i^{(\nu)}) \in \bar{\Gamma}\} && \text{in case of } T^u, \\
 \tau &= \inf \{i > 0 \mid (\pi_i^{m(\mu, \nu)}, \bar{\pi}_i^{m(\nu)}) \in \bar{\Gamma}\} && \text{in case of } \bar{T}_m^u, \\
 \tau &= \inf \{i > 0 \mid (\pi_i^{m(\mu, \nu)}, \pi_i^{m(\bar{\nu})}) \in \tilde{\Gamma}\} && \text{in case of } T_m^u
 \end{aligned}$$

and show suitable versions of Lemmas 2.1–2.3 and of Corollaries 2.2 and 2.3.

For case (v), we use the tightness and Feller property of T^u and T_m^u for the existence, while for the uniqueness we use the fact that with positive probability $r(\pi_i^{(\nu)}(\psi \wedge \psi_n))$ and $r(\pi_i^{m(\nu)}(\psi \wedge \psi_n))$ are equal zero, implying by (A3)–(A4) the uniqueness of the invariant set. The details are similar to those in the proofs of Theorem 2.3 and Proposition 3.2. \square

The next result is fundamental for our further approximations.

PROPOSITION 4.2. *Assume (A1)–(A6) together with (B1), (B4), or (C1), (C2), or (C1) and the compactness of E . Then, under (B1), (B4) for $u \in \mathcal{A}(L, n)$, and under (C1) for $u \in \mathcal{A}^r(L, n)$, we have*

$$(4.14) \quad \bar{\Psi}_m^u \Rightarrow \Psi^u \quad \text{for } m \rightarrow \infty.$$

Proof. For the case when (B1), (B4) or (C1), (C2) are satisfied, we follow the arguments of the proof of Theorem 2.2, in particular, Lemmas 2.3 and 2.4 and a suitable version of Proposition 2.2. Under (C1) and the compactness of E , by analogy to the first part of the proof of Theorem 2.1, it suffices to show that, for any $F \in C(\mathcal{P}(E) \times \mathcal{P}(E))$,

$$(4.15) \quad \bar{T}_m^u F(\mu, \nu) \rightarrow T^u F(\mu, \nu) \quad \text{for } m \rightarrow \infty$$

uniformly in $(\mu, \nu) \in \mathcal{P}(E) \times \mathcal{P}(E)$. We have

$$(4.16) \quad \begin{aligned} & |T^u F(\mu, \nu) - \bar{T}_m^u F(\mu, \nu)| \\ & \leq \int_E \int_E \int_{\mathcal{R}^d} (2\pi)^{-d/2} \exp \left[-\frac{1}{2} (y - h(z), y - h(z)) \right] \\ & \cdot |F(R(y, \zeta^u(\mu, \nu)), R(y, \zeta^u(\nu))) - F(R(y, \zeta^{\bar{\mathcal{P}}_m^u}(\mu, \nu)), \bar{M}_m^u(y, \nu))| dy \\ & \cdot P^{u(\nu)}(x, dz) \mu(dx) + \|F\| \int_E \int_{\mathcal{R}^d} \left| \int_E (2\pi)^{-d/2} \exp \left[-\frac{1}{2} (y - h(z), y - h(z)) \right] \right. \\ & \cdot (P^{u(\nu)}(x, dz) - P^{\bar{\mathcal{P}}_m^u(\nu)}(x, dz)) \left. dy \mu(dx) \right| = I_m + II_m. \end{aligned}$$

By a suitable version of Lemma A.1.2, we have that $\bar{M}_m^u(y, \nu) \Rightarrow R(y, \zeta^u(\nu))$, uniformly for $y \in B$, a compact subset of \mathcal{R}^d and $\nu \in \mathcal{P}(E)$. Moreover, by Lemma 3.1, $R(y, \zeta^{\bar{\mathcal{P}}_m^u}(\mu, \nu)) \Rightarrow R(y, \zeta^u(\mu, \nu))$ uniformly for $y \in B$, $(\nu, \mu) \in \mathcal{P}(E) \times \mathcal{P}(E)$. Therefore, repeating the considerations of the proof of Proposition 2.1, we conclude that $I_m + II_m \rightarrow 0$ uniformly in $(\nu, \mu) \in \mathcal{P}(E) \times \mathcal{P}(E)$. \square

Note that, if the state process (x_i) is governed by $P^{u(\pi_i^{(\mu)})}(\cdot, \cdot)$ with $\pi_i^{(\mu)}$ given by (4.2) for $\nu = \mu$ and starts with the initial law μ , then clearly the processes $\pi_i^{\mu, u}$ (see (1.32)), $\pi_i^{(\mu, \mu)}$ (see (4.3)) and $\pi_i^{(\mu)}$ (see (4.2)) all coincide. Given $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$, we then have, under the assumptions of Theorem 2.3,

$$(4.17) \quad \begin{aligned} J_\mu(u) &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_\mu \{c(x_i, u(\pi_i^{(\mu)}))\} \\ &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_\mu \left\{ \int_E c(z, u(\pi_i^{\mu, u})) \pi_i^{\mu, u}(dz) \right\} \\ &= \int_{\mathcal{P}(E)} \int_E c(z, u(\nu)) \nu(dz) \Phi^u(d\nu) \end{aligned}$$

valid for all μ for which (B3) holds in the cases (i) and (ii) specified in Theorem 2.3, and for all $\mu \in \mathcal{P}(E)$ in the remaining cases (iii) and (iv).

In what follows, given any sequence (η_i) of U -valued random variables, we consider the cost function $J_\mu((\eta_i))$ as given by

$$(4.18) \quad J_\mu((\eta_i)) = \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_\mu\{c(x_i, \eta_i)\}.$$

LEMMA 4.3. Assume the state process (x_i) is governed by $P^{u(\pi_i^{(\nu)})}(\cdot, \cdot)$ with $u \in \mathcal{A}(L, n)$ or $\mathcal{A}'(L, n)$ for a generic $\nu \in \mathcal{P}(E)$, but starts with the given initial law μ . Let the assumptions of Proposition 4.1(i), (ii), or (v) hold. Then for all pairs $(\mu, \nu) \in \mathcal{P}(E) \times \mathcal{P}(E)$, we have with $u_i = u(\pi_i^{(\nu)})$

$$(4.19) \quad \begin{aligned} J_\mu((u_i)) &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_\mu\{c(x_i, u(\pi_i^{(\nu)}))\} \\ &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_\mu\left\{\int_E c(z, u(\pi_i^{(\nu)})) \pi_i^{(\mu, \nu)}(dz)\right\} \\ &= \int_{\mathcal{P}(E) \times \mathcal{P}(E)} \int_E c(z, u(\nu')) \mu'(dz) \Psi^u(d\mu' \times d\nu') \\ &= \int_{\mathcal{P}(E)} \int_E c(z, u(\nu')) \nu'(dz) \Phi^u(d\nu'). \end{aligned}$$

Proof. The first identity in (4.19) follows from (4.4). By a suitable version of the second part of Theorem 2.2 in the case of assumptions (i) and (ii) of Proposition 4.1, and by the weak convergence of the Cesaro averages, the Feller property of T^u and the uniqueness of its invariant measure Ψ^u in case of assumption (v), we obtain the second identity. Since the value of the integral

$$\int_{\mathcal{P}(E) \times \mathcal{P}(E)} c(z, u(\nu')) \mu'(dz) \Psi^u(d\mu' \times d\nu')$$

does not depend on μ or $\nu \in \mathcal{P}(E)$, this is also the limit for the case where $\nu = \mu$. Then via (4.17) we obtain the last identity. \square

We can now state the main result of this section contained in Theorem 4.1 and Corollary 4.1.

THEOREM 4.1. Let $u \in \mathcal{A}(L, n)$ or $\mathcal{A}'(L, n)$ be given and define a control $u_i^{m, \bar{\nu}}$ by

$$(4.20) \quad u_i^{m, \bar{\nu}} = \tilde{\mathcal{L}}_m u(\pi_i^{m(\bar{\nu})}),$$

where $(\pi_i^{m(\bar{\nu})})$ is the approximating filtering process on the simplex S^m given by (4.6) for an arbitrary initial measure $\nu \in \mathcal{P}(E)$. Under the assumptions of Proposition 4.2, if $m > m_0(n)$, for $(\mu, \bar{\nu}) \in \mathcal{P}(E) \times S$, we have

$$(4.21) \quad \begin{aligned} J_\mu((u_i^{m, \bar{\nu}})) &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_\mu\{c(x_i, \tilde{\mathcal{L}}_m u(\pi_i^{m(\bar{\nu})}))\} \\ &= \limsup_{t \rightarrow \infty} t^{-1} \sum_{i=0}^{t-1} E_\mu\left\{\int_E c(z, \tilde{\mathcal{L}}_m u(\pi_i^{m(\bar{\nu})})) \pi_i^{(\mu, \nu)}(dz)\right\} \\ &= \int_{\mathcal{P}(E) \times S^m} \int_E c(z, \tilde{\mathcal{L}}_m u(s)) \mu'(dz) \Psi_m^u(d\mu' \times ds) \\ &= \int_{\mathcal{P}(E) \times \mathcal{P}(E)} \int_E c(z, \tilde{\mathcal{L}}_m u(\nu')) \mu'(dz) \bar{\Psi}_m^u(d\mu' \times d\nu'). \end{aligned}$$

Furthermore, we have

$$(4.22) \quad \lim_{m \rightarrow \infty} J_\mu((u_i^{m,\bar{v}})) = J_\mu((u_i)),$$

where $u_i = u(\pi_i^{\mu,u})$ is the control obtained from the given $u \in \mathcal{A}(L, n)$, evaluated at the original filtering process $(\pi_i^{\mu,u})$.

Proof. Up to the last identity, (4.21) follows from a suitable version of the second part of Theorem 2.2 in the case of assumptions (B1), (B4) or (C1), (C2), and by the weak convergence of the Cesaro averages, the Feller property of T_m^μ and the uniqueness of its invariant measure Ψ_m^μ in case of assumption (B1) and the compactness of E . The last identity in (4.21) is a consequence of Lemma 4.1 and definition (4.13) of $\tilde{\Psi}_m^\mu$. Finally, the convergence (4.22) follows from (4.21), Proposition 4.2, (4.19), and the convergence $\tilde{\mathcal{L}}_m u \rightarrow u$ for $m \rightarrow \infty$, which is uniform on compact subsets of $\mathcal{P}(E)$ (see Lemma 3.1). \square

COROLLARY 4.1. *Let u_δ^ε be the ε -optimal control function for $J_\mu^\delta(u)$ over $\mathcal{A}_m^\delta(L, n)$ or $\mathcal{A}_m^{\delta,r}(L, n)$ determined according to § 3. Let $(\pi_i^{m(\bar{\mu})})$ be the approximate filtering process on the simplex S^m , determined according to (4.6) for the given initial measure μ . Let u_i^* be the control defined by*

$$(4.23) \quad u_i^* = u_\delta^\varepsilon(\pi_i^{m(\bar{\mu})}).$$

Then, given $\varepsilon > 0$, if L, n, m are sufficiently large and δ sufficiently small, we have

$$(4.24) \quad J_\mu((u_i^*)) \leq \inf J_\mu(u) + \varepsilon,$$

where, depending on the formulation of the original problem, the infimum is taken over \mathcal{A} , $\tilde{\mathcal{A}}$, or \mathcal{A}^r ; i.e., (u_i^) is a nearly optimal control of the original problem.*

Proof. From the previous sections, we know that u_δ^ε induces a control function $\hat{u}_\delta^\varepsilon \in \mathcal{A}(L, n)$, $(\mathcal{A}^r(L, n))$ such that, defining

$$(4.25) \quad \hat{u}_i = \hat{u}_\delta^\varepsilon(\pi_i^{\mu,u}),$$

where $(\pi_i^{\mu,u})$ is the original filtering process given by (1.32), we have for sufficiently large L, n, m and small δ

$$(4.26) \quad J_\mu((\hat{u}_i)) \leq \inf J_\mu(u) + \frac{\varepsilon}{2},$$

where, again, depending on the formulation of the original problem, the infimum is taken over \mathcal{A} , $\tilde{\mathcal{A}}$, or \mathcal{A}^r . On the other hand, from Theorem 4.1, we know that, defining

$$(4.27) \quad \hat{u}_i^{m,\bar{\mu}} = \tilde{\mathcal{L}}_m \hat{u}_\delta^\varepsilon(\pi_i^{m(\bar{\mu})}),$$

where $(\pi_i^{m(\bar{\mu})})$ is the approximate filtering process on S^m obtained from (4.6) for the given initial measure μ , we have for sufficiently large m

$$(4.28) \quad J_\mu((\hat{u}_i^{m,\bar{\mu}})) \leq J_\mu((\hat{u}_i)) + \frac{\varepsilon}{2}.$$

Since

$$(4.29) \quad \tilde{\mathcal{L}}_m \hat{u}_\delta^\varepsilon = u_\delta^\varepsilon,$$

we have

$$(4.30) \quad u_i^* = \hat{u}_i^{m,\bar{\mu}},$$

and, combining (4.26), (4.28), and (4.30), we obtain the desired (4.24). \square

Appendix A.1 (Auxiliary results and proofs for § 2.1). We first recall without proof the following lemma.

LEMMA A.1.1. *Suppose that (M_1, ρ_1) , (M_2, ρ_2) are metric spaces, $F: (M_1, \rho_1) \rightarrow (M_2, \rho_2)$ is a continuous map and $K \subset M_1$ is a compact set. Moreover, $F_m: (M_1, \rho_1) \rightarrow (M_2, \rho_2)$ are measurable mappings that converge to F , uniformly on compact subsets of M_1 . Then, for any $\varepsilon > 0$, there exists $\delta > 0$ and m_0 such that for $m > m_0$*

$$(A.1.1) \quad \rho_1(x, x') < \delta \Rightarrow \rho_2(F(x), F_m(x')) < \varepsilon \quad \text{for all } x \in K, \quad x' \in M_1.$$

LEMMA A.1.2. *Under the assumptions of Theorem 2.1, we have*

$$(A.1.2) \quad M_m^{u_m}(y, \nu) \Rightarrow M^u(y, \nu) \quad \text{as } m \rightarrow \infty$$

uniformly for $y \in B$, $\nu \in \Gamma$, which are compact subsets of \mathcal{R}^d and $\mathcal{P}(E)$, respectively, and where $M^u(y, \nu)$ is given in (1.32).

Proof. Given $\phi \in C(E)$, we have

$$(A.1.3) \quad \begin{aligned} & \left| \int_E \int_E \exp \left[(y, h_m(z)) - \frac{1}{2} (h_m(z), h_m(z)) \right] \phi(z) P_m^{u_m(\nu)}(x, dz) \nu(dx) \right. \\ & \quad \left. - \int_E \int_E \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \phi(z) P^{u(\nu)}(x, dz) \nu(dx) \right| \\ & \leq \left| \int_E \int_E \left(\exp \left[(y, h_m(z)) - \frac{1}{2} (h_m(z), h_m(z)) \right] \right. \right. \\ & \quad \left. \left. - \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \right) \phi(z) P_m^{u_m(\nu)}(x, dz) \nu(dx) \right| \\ & \quad + \left| \int_E \int_E \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \phi(z) \right. \\ & \quad \left. \cdot (P_m^{u_m(\nu)}(x, dz) - P^{u(\nu)}(x, dz)) \nu(dx) \right| \\ & = I_m + II_m. \end{aligned}$$

Now

$$(A.1.4) \quad \begin{aligned} I_m & \leq \left(|y| \|h_m - h\| + \frac{1}{2} \|(h_m, h_m) - (h, h)\| \right) \\ & \quad \cdot \int_E \int_E \left(\exp \left[(y, h_m(z)) - \frac{1}{2} (h_m(z), h_m(z)) \right] \right. \\ & \quad \left. + \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \right) |\phi(z)| P_m^{u_m(\nu)}(x, dz) \nu(dx) \\ & \leq \left(|y| \|h_m - h\| + \frac{1}{2} \|(h_m, h_m) - (h, h)\| \right) (e^{|y| \|h_m\|} + e^{|y| \|h\|}) \|\phi\|, \end{aligned}$$

which by (2.3) tends to zero for $m \rightarrow \infty$, uniformly in $y \in B$.

Note now that by the Stone–Weierstrass theorem [11, Thm. 9.28 and Prob. 9.32], the function

$$(A.1.5) \quad B \times [-\|h\|, \|h\|]^d \ni (y, r) \rightarrow \exp[(y, r) - \frac{1}{2}(r, r)]$$

can be uniformly approximated with the use of functions $g_k(y, r) = \sum_{i=1}^k b_i(y) c_i(r)$,

where $b_i \in C(B)$, $c_i \in C([- \|h\|, \|h\|]^d)$; i.e., given $\varepsilon > 0$, there exists g_k such that

$$(A.1.6) \quad \sup_{y \in B} \sup_{r \in [- \|h\|, \|h\|]^d} \left| \exp \left[(y, r) - \frac{1}{2} (r, r) \right] - \sum_{i=1}^{i_k} b_i(y) c_i(r) \right| < \varepsilon.$$

Using this approximation and noting that from the compactness of Γ , for $\varepsilon > 0$ given, there is a compact set $K \subset E$ such that $\nu(K) \geq 1 - \varepsilon$ for $\nu \in \Gamma$. We then obtain

$$(A.1.7) \quad \begin{aligned} \Pi_m &\leq \left| \int_K \int_E \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \phi(z) \right. \\ &\quad \cdot (P_m^{u_m(\nu)}(x, dz) - P^{u(\nu)}(x, dz)) \nu(dx) \Big| + 2\varepsilon \|\phi\| e^{|y| \|h\|} \\ &\leq \left| \int_K \int_E \sum_{i=1}^{i_k} b_i(y) c_i(h(z)) \phi(z) (P_m^{u_m(\nu)}(x, dz) - P^{u(\nu)}(x, dz)) \nu(dx) \right| \\ &\quad + 2\varepsilon \|\phi\| + 2\varepsilon \|\phi\| e^{|y| \|h\|} \\ &\leq \sum_{i=1}^{i_k} |b_i(y)| \sup_{x \in K} \left| \int_E c_i(h(z)) \phi(z) (P_m^{u_m(\nu)}(x, dz) - P^{u(\nu)}(x, dz)) \right| \\ &\quad + 2\varepsilon \|\phi\| (1 + e^{|y| \|h\|}). \end{aligned}$$

Letting $m \rightarrow \infty$, from (2.1) and (2.2) we get

$$(A.1.8) \quad \lim_{m \rightarrow \infty} \sup_{y \in B} \Pi_m \leq 2\varepsilon \|\phi\| \left(1 + \sup_{y \in B} e^{|y| \|h\|} \right).$$

From (A.1.3), (A.1.4), (A.1.7), and (A.1.8), letting $\varepsilon \rightarrow 0$, we finally obtain (A.1.2), which finishes the proof of Lemma A.1.2. \square

Proof of Proposition 2.1. For $F \in C(\mathcal{P}(E))$, we have

$$(A.1.9) \quad \begin{aligned} &|\Pi_m^{u_m}(\nu, F_m) - \Pi^u(\nu, F)| \\ &= \left| (2\pi)^{-(d/2)} \int_E \int_{\mathcal{R}^d} \left[\int_E \exp \left[-\frac{1}{2} (y - h_m(\eta), y - h_m(\eta)) \right] \right. \right. \\ &\quad \cdot P_m^{u_m(\nu)}(\zeta, d\eta) F_m(M_m^{u_m}(y, \nu)) \\ &\quad \left. \left. - \int_E \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] P^{u(\nu)}(\zeta, d\eta) F(M^u(y, \nu)) \right] dy \nu(d\zeta) \right| \\ &\leq \left| (2\pi)^{-(d/2)} \int_E \int_{\mathcal{R}^d} \int_E \left(\exp \left[-\frac{1}{2} (y - h_m(\eta), y - h_m(\eta)) \right] \right. \right. \\ &\quad \left. \left. - \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] \right) \right. \\ &\quad \cdot P_m^{u_m(\nu)}(\zeta, d\eta) F_m(M_m^{u_m}(y, \nu)) dy \nu(d\zeta) \Big| \\ &\quad + \left| (2\pi)^{-(d/2)} \int_E \int_{\mathcal{R}^d} \int_E \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] \right. \\ &\quad \cdot (P_m^{u_m(\nu)}(\zeta, d\eta) - P^{u(\nu)}(\zeta, d\eta)) F_m(M_m^{u_m}(y, \nu)) dy \nu(d\zeta) \Big| \\ &\quad + \left| (2\pi)^{-(d/2)} \int_E \int_{\mathcal{R}^d} \int_E \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] P^{u(\nu)}(\zeta, d\eta) \right. \\ &\quad \left. \cdot (F_m(M_m^{u_m}(y, \nu)) - F(M^u(y, \nu))) dy \nu(d\zeta) \right| = \text{I}_m + \text{II}_m + \text{III}_m. \end{aligned}$$

We now obtain

$$\begin{aligned}
 I_m &\leq (2\pi)^{-(d/2)} \int_E \int_{\mathcal{R}^d} \int_E (|y| \|h_m - h\| + \|(h_m, h_m) - (h, h)\|) \\
 &\quad \cdot \left(\exp \left[-\frac{1}{2} (y - h_m(\eta), y - h_m(\eta)) \right] + \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] \right) \\
 &\quad \cdot b_F P_m^{u(\nu)}(\zeta, d\eta) dy \nu(d\zeta) \\
 &\leq \int_E \int_E \left\{ (2\pi)^{-(d/2)} \int_{\mathcal{R}^d} \left(\exp \left[-\frac{1}{2} (y - h_m(\eta), y - h_m(\eta)) \right] \right. \right. \\
 (A.1.10) \quad &\quad \left. \left. + \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] \right) \|(h_m, h_m) - (h, h)\| dy \right. \\
 &\quad \left. + \left[\left((2\pi)^{-(d/2)} \int_{\mathcal{R}^d} (y, y) \exp \left[-\frac{1}{2} (y - h_m(\eta), y - h_m(\eta)) \right] dy \right)^{1/2} \right. \right. \\
 &\quad \left. \left. + \left((2\pi)^{-(d/2)} \int_{\mathcal{R}^d} (y, y) \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] dy \right)^{1/2} \right] \right. \\
 &\quad \left. \cdot \|h_m - h\| \right\} b_F P_m^{u(\nu)}(\zeta, d\eta) \nu(d\zeta) \\
 &\leq 2 \|(h_m, h_m) - (h, h)\| b_F + (2 + \|(h_m, h_m)\| + \|(h, h)\|) \|h_m - h\| b_F \rightarrow 0 \\
 &\quad \text{for } m \rightarrow \infty.
 \end{aligned}$$

To evaluate II_m and III_m , let $\Gamma \subset \mathcal{P}(E)$ be a compact set. For any $\varepsilon > 0$, there exist compact sets $K \subset E$, $B \subset \mathcal{R}^d$ such that for $\nu \in \Gamma$, $\nu(K) \geq 1 - \varepsilon$ and

$$(A.1.11) \quad (2\pi)^{-(d/2)} \inf_{\eta \in E} \int_B \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] dy \geq 1 - \varepsilon.$$

By Proposition 1.2(i), $M^u(B, \Gamma)$ is a compact subset of $\mathcal{P}(E)$. Applying Lemma A.1.1 to $(M_1, \rho_1) = (\mathcal{P}(E), \rho_w)$, $(M_2, \rho_2) = (\mathcal{R}, |\cdot|)$, $K = M(B, \Gamma)$, where ρ_w stands for the metric compatible with weak convergence on $\mathcal{P}(E)$, for a sufficiently large m , by Lemma A.1.2 we have

$$(A.1.12) \quad \sup_{y \in B} \sup_{\nu \in \Gamma} |F_m(M_m^{u_m}(y, \nu)) - F(M^u(y, \nu))| \leq \varepsilon.$$

Estimating II_m , we again use the approximation (A.1.6), obtaining

$$\begin{aligned}
 (A.1.13) \quad II_m &\leq \left| (2\pi)^{-(d/2)} \int_K \int_B \int_E \exp \left[-\frac{1}{2} (y - h(\eta), y - h(\eta)) \right] \right. \\
 &\quad \cdot (P_m^{u_m(\nu)}(\zeta, d\eta) - P^{u(\nu)}(\zeta, d\eta)) F_m(M_m^{u_m}(y, \nu)) dy \nu(d\zeta) \left. \right| + 2b_F \varepsilon + 2b_F \varepsilon \\
 &\leq 6b_F \varepsilon + (2\pi)^{-(d/2)} \int_K \int_B \exp \left[-\frac{1}{2} (y, y) \right] \sum_{i=1}^{i_k} |c_i(y)|
 \end{aligned}$$

$$\begin{aligned}
& \cdot \left| \int_E b_i(h(\eta))(P_m^{u_m(\nu)}(\zeta, d\eta) - P^{u(\nu)}(\zeta, d\eta)) \right| \cdot |F_m(M_m^u(y, \nu))| dy \nu(d\zeta) \\
& \leq 6b_F \varepsilon + b_F (2\pi)^{-(d/2)} \sum_{i=1}^{i_k} \int_B \exp \left[-\frac{1}{2}(y, y) \right] |c_i(y)| dy \\
& \cdot \sup_{\zeta \in K} \left| \int_E b_i(h(\eta))(P_m^{u_m(\nu)}(\zeta, d\eta) - P^{u(\nu)}(\zeta, d\eta)) \right|.
\end{aligned}$$

Then, using (2.1) and (2.2), we get for $\nu \in \Gamma$

$$(A.1.14) \quad \limsup_{m \rightarrow \infty} \Pi_m \leq 6b_F \varepsilon.$$

Finally, by (A.1.12) we have for sufficiently large m

$$\begin{aligned}
(A.1.15) \quad \text{III}_m & \leq (2\pi)^{-(d/2)} \int_K \int_B \int_E \exp \left[-\frac{1}{2}(y - h(\eta), y - h(\eta)) \right] P^{u(\nu)}(\zeta, d\eta) \\
& \cdot |F_m(M_m^u(y, \nu)) - F(M^u(y, \nu))| dy \nu(d\zeta) + 2b_F \varepsilon + 2b_F \varepsilon \\
& \leq \sup_{y \in B} \sup_{\nu \in \Gamma} |F_m(M_m^u(y, \nu)) - F(M^u(y, \nu))| \\
& \cdot (2\pi)^{-(d/2)} \int_K \int_E \int_B \exp \left[-\frac{1}{2}(y - h(\eta), y - h(\eta)) \right] \\
& \cdot dy P^{u(\nu)}(\zeta, d\eta) \nu(d\zeta) + 4b_F \varepsilon \\
& \leq \varepsilon(4b_F + 1).
\end{aligned}$$

In summary, from (A.1.9), (A.1.10), (A.1.14), (A.1.15),

$$(A.1.16) \quad \limsup_{m \rightarrow \infty} \sup_{\nu \in \Gamma} \left| \Pi_m^u(\nu, F) - \Pi^u(\nu, F) \right| \leq \varepsilon(10b_F + 1).$$

Letting $\varepsilon \rightarrow 0$, we obtain the proof of Proposition 2.1. \square

Proof of Lemma 2.1. Let us first consider the case (B1), (B4) and, to be specific, assume that alternative (1.8) holds. Given n , we can then find $n_1 > n$ such that $h^j(z) \geq \|h^j\|_n + \varepsilon$ for some $\varepsilon > 0$ and all $z \notin K_{n_1}$, where $\|h^j\|_n = \sup_{x \in K_n} |h^j(x)|$. Then, for $y^j > y_0 > 0$ and $|y^i| \leq M$, if $i \neq j$,

$$\begin{aligned}
R(y, \zeta^u(\nu))(K_{n_1}) & \leq e^{2\|h\|M} \\
& \cdot \left(\int_{K_{n_1}^c} \exp \left[(y, h(z) - \|h^j\|_n) - \frac{1}{2}(h(z), h(z)) \right] \zeta^u(\nu)(dz) \right)^{-1} \\
& \leq e^{2\|h\|M} e^{(1/2)\|h\|^2} e^{-\varepsilon y_0} e^{2\|h\|M} \frac{1}{\zeta^u(\nu)(K_{n_1}^c)} \\
& \leq e^{4\|h\|M} e^{(1/2)\|h\|^2} \frac{1}{\alpha} e^{-\varepsilon y_0} \leq 1 - \gamma.
\end{aligned}$$

By (2.18), this then implies the first statement. Similarly, we obtain $M_m^u(y, \nu)(K_n^c) \geq \gamma$ where m_0 is such that, for $m > m_0$, $h_m^j(z) \geq \|h_m^j\| + \varepsilon$. Analogously, we may study the case (C1), (C2). \square

Proof of Lemma 2.4. We prove case (i) of Lemma 2.4 for alternative (1.8). If $P\{M^u(Z, \nu)(\phi) = b\} > 0$, then by continuity (see Proposition 1.2), $M^u(Z, \nu)(\phi) = b$ for all y from some open subset G of \mathbb{R}^d . By (A3), it then follows that $b > 0$, since, for nonnegative $\phi(z) \neq 0$,

$$\int_E \int_E \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \phi(z) P^{u(\nu)}(x, dz) \nu(dx) > 0.$$

Note now that

$$\int_E \int_E \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] (\phi(z) - b) P^u(x, dz) \nu(dx) = 0$$

for $y \in G$. Differentiating m times with respect to y_j , we obtain

$$\int_E \int_E (h^j(z))^m \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] (\phi(z) - b) P^u(x, dz) \nu(dx) = 0$$

for $y \in G$, $m = 0, 1, 2, \dots$. Therefore, for any continuous function $c : [-\|h\|, \|h\|] \rightarrow \mathbb{R}$,

$$\int_E \int_E c(h^j(z)) \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] (\phi(z) - b) P^u(x, dz) \nu(dx) = 0$$

for $y \in G$. There exists a compact set K with $K_n \subset K$ such that

$$\sup_{z \in E} h^j(z) \geq \inf_{z \in K^c} h^j(z) = a_1 > \sup_{z \in K_n} h^j(z).$$

Let g_1 be a strictly increasing, continuous, and bounded function: $[-\|h\|, \|h\|] \rightarrow \mathbb{R}$ and set

$$g_2(a) = \begin{cases} g_1(a) & \text{for } a < a_1, \\ g_1(a_1) & \text{for } a > a_1. \end{cases}$$

Then, for $y \in G$,

$$\begin{aligned} & \int_E \int_E (g_1 - g_2)(h^j(z)) \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] \\ & \cdot (\phi(z) - b) P^u(x, dz) \nu(dx) = 0 \end{aligned}$$

and, consequently,

$$\int_E \int_{K_n^c} (g_1 - g_2)(h^j(z)) \exp \left[(y, h(z)) - \frac{1}{2} (h(z), h(z)) \right] (-b) P^u(x, dz) \nu(dx) = 0$$

for $y \in G$. However, $(g_1 - g_2)(h^j(z))$ is strictly positive on an open subset of K_n^c . Therefore, by (A3), $b = 0$, implying thus a contradiction.

Case (ii) can be shown in a similar way. \square

LEMMA A.1.3. *Consider the assumptions of Proposition 2.2 and suppose further that there exist uniformly bounded sequences $F_i^m \in \mathcal{B}(\mathcal{P}(E))$ ($i = 1, 2, \dots, q$) such that $F_i^m \rightarrow F_i$ for $m \rightarrow \infty$, uniformly on compact subsets of $\mathcal{P}(E)$. Then, if $\nu_m \Rightarrow \nu$, we have for each positive integer q*

$$\begin{aligned} (A.1.17) \quad & E_{\nu_m}^u \{ F_1^m(\pi_1^{m, \nu_m, u_m}) \cdot \dots \cdot F_q^m(\pi_q^{m, \nu_m, u_m}) \} \\ & \rightarrow E_{\nu}^u \{ F_1(\pi_1^{\nu, u}) \cdot \dots \cdot F_q(\pi_q^{\nu, u}) \}. \end{aligned}$$

Proof. We prove by induction. Step $q = 1$ follows from Corollary 2.1. Assume that (A.1.17) holds for q ; then for $q + 1$ we have

$$(A.1.18) \quad \begin{aligned} E_{\nu_m}^{u_m} \{ F_1^m(\pi_1^{m,\nu_m,u_m}) \cdots F_q^m(\pi_q^{m,\nu_m,u_m}) F_{q+1}^m(\pi_{q+1}^{m,\nu_m,u_m}) \} \\ = E_{\nu_m}^{u_m} \{ F_1^m(\pi_1^{m,\nu_m,u_m}) \cdots F_q^m(\pi_q^{m,\nu_m,u_m}) \Pi_m^u(\pi_q^{m,\nu_m,u_m}, F_{q+1}^m) \}. \end{aligned}$$

Recall now from Proposition 2.1 that

$$\Pi_m^u(\nu, F_{q+1}^m) \rightarrow \Pi^u(\nu, F_{q+1}) \quad \text{for } m \rightarrow \infty$$

uniformly on compact subsets of $\mathcal{P}(E)$. By the induction hypothesis, we then have the convergence of (A.1.18) to

$$\begin{aligned} E_{\nu}^u \{ F_1(\pi_1^{\nu,u}) \cdots F_q(\pi_q^{\nu,u}) \Pi^u(\pi_q^{\nu,u}, F_{q+1}) \} \\ = E_{\nu}^u \{ F_1(\pi_1^{\nu,u}) \cdots F_{q+1}(\pi_{q+1}^{\nu,u}) \}. \end{aligned} \quad \square$$

Appendix A.2 (Auxiliary results and proofs for § 3.1).

LEMMA A.2.1. *Let $u \in \mathcal{A}(L, n)$ or $\mathcal{A}^r(L, n)$. Given the partially observed system (CS_1) , let*

$$(A.2.1) \quad x_i^m = \sum_{k=1}^{k_m} k \chi_{B_k^m}(\bar{x}_i^m).$$

Then x_i^m is a particular version of the state process in the system (CS_2) corresponding to the transition matrix $P_m^{\tilde{\mathcal{L}}_m^u}(k, p)$. Furthermore, the observations \bar{y}_i in (3.16) and (3.18) coincide and

$$(A.2.2) \quad \bar{\pi}_i^{m,\mu,u}(B_k^m) = \pi_i^{m,\bar{\mu},\tilde{\mathcal{L}}_m^u}(k), \quad k = 1, 2, \dots, k_m;$$

i.e., $\pi_i^{m,\bar{\mu},\tilde{\mathcal{L}}_m^u}$ is a restriction of $\bar{\pi}_i^{m,\mu,u}$ to (S^m) . In addition, if $F: \mathcal{R}^{k_m} \rightarrow \mathcal{R}$ is bounded measurable, then

$$(A.2.3) \quad \begin{aligned} \int_{\mathcal{P}(E)} F(\nu(B_1^m), \dots, \nu(B_{k_m}^m)) \bar{\Pi}_m^u(\nu', d\nu) \\ = \int_{S^m} F(s_1, \dots, s_{k_m}) \Pi_m^{\tilde{\mathcal{L}}_m^u}(\nu'(B_1^m), \dots, \nu'(B_{k_m}^m), ds). \end{aligned}$$

Proof. By the definition of h_m in (3.1) and by (A.2.1), we clearly have that the observation processes \bar{y}_i of the types (3.16) and (3.18) coincide. Furthermore, the initial law for (x_i^m) is $\bar{\mu}$, which, by (3.14) and (3.15) as well as by (3.8), implies that the transition matrix for (x_i^m) at the initial stage $i = 0$ is $P_m^{\tilde{\mathcal{L}}_m^u(\bar{\mu})}(x_0^m, p)$. The fact that the transition law for x_i^m is $P_m^{\tilde{\mathcal{L}}_m^u}(k, p)$ for any period i follows then by induction using always (3.14), (3.15), and (3.8). Relation (A.2.2) now follows, again by induction, comparing the values of $\bar{M}_m^u(\bar{y}_{i+1}, \bar{\pi}_i^{m,\mu,u})$ and $\bar{M}_m^{\tilde{\mathcal{L}}_m^u}(\bar{y}_{i+1}, \pi_i^{m,\bar{\mu},\tilde{\mathcal{L}}_m^u})$.

Coming to the second part of the lemma, let $G \in \mathcal{B}(S)$ and $F = 1_G$. Then

$$\begin{aligned} \int_{\mathcal{P}(E)} 1_G(\nu(B_1^m), \dots, \nu(B_{k_m}^m)) \bar{\Pi}_m^u(\nu', d\nu) &= \bar{\Pi}_m^u(\nu', D) \\ &= \Pi_m^{\tilde{\mathcal{L}}_m^u}(\nu'(B_1^m), \dots, \nu'(B_{k_m}^m), G), \end{aligned}$$

where $D = \{\nu \in \mathcal{P}(E) \mid (\nu(B_1^m), \dots, \nu(B_{k_m}^m)) \in G\}$. Thus (A.2.3) is satisfied for simple functions and, consequently, by a standard procedure also for any bounded measurable function F . \square

Proof of Proposition 3.1. Given $A \in \mathcal{B}(\mathcal{P}(E))$, $u \in \mathcal{A}(L, n)$, or $\mathcal{A}^r(L, n)$, note from (3.14), (3.19), (3.22) that the value of $\bar{\Pi}_m^u(\nu, A)$ depends only on $(\nu(B_1^m), \dots, \nu(B_{k_m}^m))$, i.e., there is a measurable, bounded function F such that

$$(A.2.4) \quad \bar{\Pi}_m^u(\nu, A) = F(\nu(B_1^m), \dots, \nu(B_{k_m}^m)).$$

Therefore, using (3.25), (A.2.4), and (A.2.3), we obtain

$$(A.2.5) \quad \begin{aligned} & \int_{\mathcal{P}(E)} \bar{\Pi}_m^u(\nu, A) \bar{\Phi}_m^u(d\nu) \\ &= \int_{\mathcal{P}(E)} \bar{\Pi}_m^u(\nu, A) \int_{S^m} \bar{\Pi}_m^u\left(\sum_{k=1}^{k_m} s_k \delta_{z_k^m}, d\nu\right) \Phi_m^{\tilde{\mathcal{L}}_m^u}(ds) \\ &= \int_{S^m} \left(\int_{\mathcal{P}(E)} \bar{\Pi}_m^u(\nu, A) \bar{\Pi}_m^u\left(\sum_{k=1}^{k_m} s_k \delta_{z_k^m}, d\nu\right) \right) \Phi_m^{\tilde{\mathcal{L}}_m^u}(ds) \\ &= \int_{S^m} \int_{S^m} \bar{\Pi}_m^u\left(\sum_{k=1}^{k_m} a_k \delta_{z_k^m}, A\right) \Pi_m^{\tilde{\mathcal{L}}_m^u}(s_1, \dots, s_{k_m}, da) \Phi_m^{\tilde{\mathcal{L}}_m^u}(ds) \\ &= \int_{S^m} \bar{\Pi}_m^u\left(\sum_{k=1}^{k_m} a_k \delta_{z_k^m}, A\right) \Phi_m^{\tilde{\mathcal{L}}_m^u}(da) = \bar{\Phi}_m^u(A) \end{aligned}$$

with $a = (a_1, \dots, a_{k_m}) \in S^m$, which is our claim. \square

Proof of Proposition 3.2. By the properties (D_m) and (A5), there exists, both under (B1) and (C1), an $m_0(n)$ with the properties formulated in the first part of Proposition 3.2. The proof of the uniqueness of the invariant measure Φ_m^u is based on arguments similar to those used in the proof of Theorem 2.3. Namely, proceeding analogously to (2.35)–(2.37), we show that there exists a unique invariant set for the controlled filtering process, which implies the uniqueness of the invariant measure Φ_m^u .

The first part of identity (3.29) now follows from the Feller property of $(\pi_i^{m, \bar{\mu}, \tilde{\mathcal{L}}_m^u})$ and the uniqueness of the invariant measure $\Phi_m^{\tilde{\mathcal{L}}_m^u}$. It remains to prove the second part of (3.29). By (3.25) we have

$$(A.2.6) \quad \begin{aligned} & \int_{\mathcal{P}(E)} \int_E c_m(x, \tilde{\mathcal{L}}_m u(\nu)) \nu(dx) \bar{\Phi}_m^u(d\nu) \\ &= \int_{\mathcal{P}(E)} \int_{S^m} \int_E c_m(x, \tilde{\mathcal{L}}_m u(\nu)) \nu(dx) \bar{\Pi}_m^u\left(\sum_{k=1}^{k_m} s_k \delta_{z_k^m}, d\nu\right) \Phi_m^{\tilde{\mathcal{L}}_m^u}(ds) \\ &= \int_{S^m} \int_{\mathcal{P}(E)} \int_E c_m(x, \tilde{\mathcal{L}}_m u(\nu)) \nu(dx) \bar{\Pi}_m^u\left(\sum_{k=1}^{k_m} s_k \delta_{z_k^m}, d\nu\right) \Phi_m^{\tilde{\mathcal{L}}_m^u}(ds). \end{aligned}$$

By (A5) and the definition of $\tilde{\mathcal{L}}_m$,

$$(A.2.7) \quad \int_E c_m(x, \tilde{\mathcal{L}}_m u(\nu)) \nu(dx) = \sum_{k=1}^{k_m} c_k^m\left(u\left(\sum_{p=1}^{k_m} \nu(B_p^m) \delta_{z_p^m}\right)\right) \nu(B_k^m)$$

is a function of $(\nu(B_1^m), \dots, \nu(B_{k_m}^m))$ only. Therefore, applying (A.2.3), we continue (A.2.6), obtaining

$$(A.2.8) \quad \begin{aligned} & \int_{\mathcal{P}(E)} \int_E c_m(x, \tilde{\mathcal{L}}_m u(\nu)) \nu(dx) \bar{\Phi}_m^u(d\nu) \\ &= \int_{S^m} \int_{S^m} \sum_{k=1}^{k_m} c_k^m\left(u\left(\sum_{p=1}^{k_m} s_p \delta_{z_p^m}\right)\right) s_k \Pi_m^{\tilde{\mathcal{L}}_m^u}(a_1, \dots, a_{k_m}, ds) \Phi_m^{\tilde{\mathcal{L}}_m^u}(da) \\ &= \int_{S^m} \sum_{k=1}^{k_m} c_k^m(\tilde{\mathcal{L}}_m u(s)) s_k \Phi_m^{\tilde{\mathcal{L}}_m^u}(ds), \end{aligned}$$

where the last identity follows from the fact that $\Phi_m^{\mathcal{F}_m^u}$ is an $\Pi_m^{\mathcal{F}_m^u}$ invariant measure. \square

REFERENCES

- [1] A. ARAPOSTATHIS, E. FERNANDEZ-GAUCHERAND, AND S. I. MARCUS, *Analysis of an adaptive control scheme for a partially observed controlled Markov chain*, in Proc. 29th CDC, Honolulu, Hawaii, 1990, pp. 1438–1444.
- [2] A. BENSOUSSAN AND W. Runggaldier, *An approximation method for stochastic control problems with partial observation of the state—A method for constructing ϵ -optimal controls*, Acta Appl. Math., 10 (1987), pp. 145–170.
- [3] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
- [4] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [5] V. S. BORKAR, *The probabilistic structure of controlled diffusion processes*, Acta Appl. Math., 11 (1988), pp. 19–48.
- [6] G. B. DI MASI AND L. STETTNER, *On adaptive control of a partially observed Markov chain*, preprint.
- [7] T. DUNCAN, B. PASIK-DUNCAN, AND L. STETTNER, *On ergodic and adaptive control problems for stochastic differential delay equations*, preprint.
- [8] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
- [9] ———, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048.
- [10] H. J. KUSHNER AND W. Runggaldier, *Nearly optimal feedback controls for stochastic systems with wideband noise disturbances*, SIAM J. Control Optim., 25 (1987), pp. 289–315.
- [11] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1968.
- [12] W. J. Runggaldier AND L. STETTNER, *On the Construction of Nearly Optimal Strategies for a General Problem of Control of Partially Observed Diffusions*, IMPAN Preprint 450, Warsaw, 1989; Stochastics Stochastics Reports, 37 (1991), pp. 15–47.
- [13] ———, *Nearly optimal controls for stochastic ergodic problems with partial observation*, IMPAN Preprint 484, Warsaw, 1991.
- [14] ———, *Partially observable control problems with compulsory shifts of the state*, IIASA Working Paper, WP-92-34, Vienna, Austria, May 1992.
- [15] A. W. SKOROKHOD, *Asymptotic Methods in the Theory of Stochastic Differential Equations*, Naukova Dumka, Kiev, 1987. (In Russian); English translation, Translated Math. Monographs, Vol. 78, American Mathematical Society, Providence, RI, 1989.
- [16] L. STETTNER, *On invariant measures of filtering processes*, in Proc. 4th Bad Honnef Conference on Stochastic Differential Systems, Lecture Notes in Control and Inform. Sci. 126, Springer-Verlag, New York, 1989, pp. 279–292.

A SUBSPACE DECOMPOSITION PRINCIPLE FOR SCALED GRADIENT PROJECTION METHODS: LOCAL THEORY*

J. C. DUNN†

Abstract. This paper is a sequel to an earlier work on a modification of the Gafni–Bertsekas scaled gradient projection method. Global convergence theorems proved in the first paper are complemented here by a local convergence analysis for the general SGP schemes and for special SGP algorithms with Newtonian scaling operators.

Key words. constrained minimization, gradient projection, Newtonian scaling, local convergence

AMS(MOS) subject classifications. 49D07, 65K10, 65B99

1. Introduction. The Goldstein–Levitin–Poljak unscaled gradient projection (GP) scheme [1], [2] inherits the basic convergence behavior of its steepest descent counterpart for unconstrained minimization [3]–[11], and costs little more to implement in the simple polyhedral and nonpolyhedral convex feasible sets commonly found in control problems, network flow problems, and other applications (e.g., Cartesian products of orthants, boxes, simplices, balls, cones, etc., [3], [5], [12]–[15]). For such problems, the GP method will compute well-conditioned nonsingular minimizers very efficiently. On the other hand, like steepest descent, the GP algorithm converges slowly to minimizers that are only moderately ill-conditioned, and some sort of Newtonian scaling of the objective function gradient is needed to restore acceptable rates of convergence in such cases. Bertsekas explains how this scaling might be done in orthants, boxes, and simplices [14], proves global and local convergence theorems in this setting, demonstrates the effectiveness of the scaled gradient projection (SGP) method for multistage bounded input optimal control problems, and proposes an extension of the scaling principle for general polyhedral feasible sets in \mathbb{R}^n . In [15], Gafni and Bertsekas describe a modified scaling principle that can be implemented in any closed convex set, prove a global convergence result at this level of generality, and show that in polyhedral sets in \mathbb{R}^n , any iterate sequence $\{u^i\}$ with a nondegenerate proper (i.e., strict) local minimizer \bar{u} in its limit point set must converge to \bar{u} , and converge superlinearly when \bar{u} is nonsingular and the objective function gradient is properly scaled.

While the global convergence theorem in [15] holds in arbitrary closed convex sets in a Hilbert space, the accompanying local superlinear convergence result is essentially limited to polyhedra since the scaling subspaces produced by the Gafni–Bertsekas dual cone decomposition in nonpolyhedral sets are too small to support Newtonian scaling operators likely to induce superlinear convergence [16]; moreover, the local result *assumes* that $\{u^i\}$ has a nonsingular minimizer \bar{u} in its limit point set, and is therefore weaker than the true local convergence theorems in [14] and [5], [8], [9] (which establish convergence of $\{u^i\}$ to *any* nonsingular local minimizer \bar{u} from all

*Received by the editors September 10, 1990; accepted for publication (in revised form) October 24, 1991. This research was supported by National Science Foundation grant DMS8702929.

†Mathematics Department, Box 8205, North Carolina State University, Raleigh, North Carolina 27695-8205.

sufficiently nearby starting points u^1). In addition, the local convergence proof in [15] is not valid in infinite-dimensional spaces. With these observations in mind, [13] and [17] consider a variant of the scaling principles in [14], [15] designed for closed convex feasible sets prescribed by finitely many smooth inequality constraints in a real Hilbert space. The new subspace decomposition and scaling principle preserves the descent and limit point stationarity properties of the algorithms in [14], [15], produces scaling subspaces that support the Newtonian scaling operators in [16], generates the same iterates as Bertsekas's original SGP algorithm near nondegenerate stationary points in orthants and boxes, is simpler than the Gafni–Bertsekas dual cone construction, and has some cost advantage over both of the earlier SGP schemes even in rudimentary polyhedral sets in \mathbb{R}^n (e.g., in a n -dimensional orthant, the derivative sign tests in [14] and cone projections in [15] can require $O(n)$ floating point comparisons beyond the $O(n)$ flops count for subspace decomposition).

The present article complements the global analysis in [17] by resolving several basic local convergence questions for the subspace decomposition principle in convex feasible sets prescribed by finitely many smooth inequality constraints, and for a generalization of the Gafni–Bertsekas dual cone decomposition principle in polyhedra; as in [17], the setting is a general real Hilbert space and the results obtained apply to finite- and infinite-dimensional nonlinear programs. For convenience, the SGP algorithms, constraint qualification, and several definitions from [8], [17] are fully described in §2, along with the Newtonian scaling operators in [16]. Section 3 then addresses the local convergence behavior of the SGP algorithms for general (i.e., not necessarily Newtonian) scaling operators that satisfy only minimal boundedness and coercivity conditions. It is first shown that the SGP sequences $\{u^i\}$ converge to a stationary point \bar{u} from all nearby starting points u^1 *only if* \bar{u} is a proper local minimizer and an isolated stationary point. A partial converse of this result is then proved for nondegenerate stationary points in \mathbb{R}^n , and several other partial converses are established in arbitrary real Hilbert spaces. The proof strategy employed here subsumes the local convergence proof in [14] and parallels the analysis in [8] and [9]. Nondegeneracy and local uniform growth conditions are imposed on the objective function near \bar{u} to insure that the iterates u^i will remain in any arbitrarily small neighborhood of \bar{u} in the feasible set if u^1 is sufficiently close to \bar{u} (i.e., to make \bar{u} a “stable” fixed point for the SGP algorithms). Additional local growth conditions on some measure of nonstationarity near \bar{u} can then be invoked to force convergence of $\{u^i\}$ to \bar{u} for all sequences $\{u^i\}$ confined to a sufficiently small neighborhood of \bar{u} (i.e., to make \bar{u} an “asymptotically stable” fixed point, or “stable local attractor”). In particular, if the required growth conditions hold and the objective and constraint function gradients are Lipschitz continuous near \bar{u} , then SGP iterates that begin near \bar{u} are shown to eventually enter and remain within the smooth manifold defined by the active inequality constraints at \bar{u} , and then converge to \bar{u} . All of the aforementioned nondegeneracy, growth, and continuity hypotheses are demonstrated for nonsingular local minimizers, and linear convergence rates are also proved in this special case; however, local asymptotic stability is established here for a much larger class of strict local minimizers, and the proof technique in [8] will yield a hierarchy of associated convergence rates for SGP processes in polyhedral sets. Thus, the analysis in §3 not only proves desirable generic local convergence properties for the new SGP subspace decomposition scheme in nonpolyhedral sets, but also substantially extends and strengthens the local convergence theorems of [14], [15] in polyhedra. In the concluding §4, the focus is narrowed to SGP algorithms that use the largest admissible scaling subspaces and the associated New-

tonian scaling operators in [16]. In this setting, the subspace decomposition scheme and the Gafni–Bertsekas dual cone decomposition scheme are shown to produce identical iterates near nondegenerate stationary points (and in particular near nonsingular minimizers) in polyhedra, and a new local superlinear convergence theorem is proved for the subspace decomposition principle in nonpolyhedral sets.

Several interesting questions are left unanswered in this investigation. For instance, while all of the SGP schemes considered here have the same basic descent and limit point stationarity properties and exhibit identical local convergence behavior near nondegenerate stationary points wherever they are jointly applicable, this does not rule out potentially significant differences in their global convergence properties (and hence their net computational costs) within specific problem classes; existing global convergence theories and published numerical experiments are simply not capable of resolving this issue at present. In addition, counterparts of the important quasi-Newton scaling operator recursions for unconstrained minimization have also received no explicit consideration here. It is easily seen that quasi-Newtonian versions of the subject SGP algorithms will retain the local superlinear convergence property if their scaling operator sequences obey asymptotic quasi-Newton conditions like (43) in the SGP framework. If the unconstrained quasi-Newton theory is a reliable guide here, then the general linear convergence rate estimate and active constraint identification results in Theorem 4 of §3 should supply the essential first steps in a demonstration of this sort.

2. SGP methods. Let Ω be a nonempty closed convex set in a real Hilbert space \mathcal{U} , and let J be a continuously Fréchet differentiable real function to be minimized over Ω . The SGP algorithms in [13]–[15], [17] generate successive feasible approximations $u^i \in \Omega$ recursively with

$$(1a) \quad u^{i+1} = P_{\Omega}(u^i + \sigma^i v^i)$$

where at each $u \in \Omega$,

$$(1b) \quad v = v_N + v_T,$$

$$(1c) \quad v_N = s_N P_N(-\nabla J(u)),$$

$$(1d) \quad v_T = P_T S_T P_{N^*}(-\nabla J(u)),$$

$$(1e) \quad N = \text{a closed convex cone containing } K(u),$$

$K(u)$ = the normal cone at u

$$(1f) \quad = \{w : \forall v \in \Omega, \langle w, v - u \rangle \leq 0\},$$

N^* = the dual cone for N

$$(1g) \quad = \{w^* : \forall w \in N, \langle w^*, w \rangle \leq 0\},$$

$$(1h) \quad T = \{v_N\}^{\perp} \cap N^*,$$

$$(1i) \quad [T] = \text{the closed linear hull of } T,$$

$$(1j) \quad \mu_3 \geq s_N \geq \mu_2 > 0,$$

S_T = a bounded linear map from $[T]$ into $[T]$ such that

$$(1k) \quad \langle P_{N^*}(-\nabla J(u)), S_T P_{N^*}(-\nabla J(u)) \rangle \geq \mu_0 \|P_{N^*}(-\nabla J(u))\|^2,$$

and

$$\begin{aligned}
 (11) \quad & \|S_T\| \leq \mu_1, \\
 & \sigma = \max s, \text{ subject to} \\
 & \frac{s}{\alpha} \in \{1, \beta, \beta^2, \dots\},
 \end{aligned}$$

and

$$\begin{aligned}
 (1m) \quad & J(u) - J(P_\Omega(u + sv)) \geq \delta \{(s_N s)^{-1} \|u + sv_T - P_\Omega(u + sv)\|^2 \\
 & + \langle P_{N^*}(-\nabla J(u)), v_T \rangle s\},
 \end{aligned}$$

and where δ and β are fixed real numbers in $(0, 1)$, α and μ_0, \dots, μ_3 are fixed positive real numbers, and P_A denotes projection into the set A .

Further restrictions are needed on the cones N in (1) to insure desirable global and local convergence behavior for the associated sequences $\{u^i\}$. The extended Gafni–Bertsekas construction treated in [17] begins with an explicit affine inequality representation

$$(2a) \quad \Omega = \{u \in \mathcal{U} : \forall j \in \mathcal{J}, \langle a^j, u \rangle - bj \leq 0\}$$

for the closed convex set Ω , where the index set \mathcal{J} is finite for polyhedral Ω , and infinite otherwise. Given an arbitrary but fixed positive number ϵ_o , we then put

$$\begin{aligned}
 (2b) \quad & d(u) = \|u - P_\Omega(u - \nabla J(u))\|, \\
 (2c) \quad & \epsilon(u) = \min\{\epsilon_o, d(u)\}, \\
 (2d) \quad & \mathcal{J}(u) = \{j \in \mathcal{J} : \langle a^j, u \rangle - bj \geq -\|a^j\|\epsilon(u)\}, \\
 (2e) \quad & C(u) = \{w : \forall j \in \mathcal{J}(u), \langle w, a^j \rangle \leq 0\},
 \end{aligned}$$

and require that

$$(2f) \quad N \supset C(u)^* = \{w^* : \forall w \in C(u), \langle w^*, w \rangle \leq 0\}$$

at each $u \in \Omega$. This restriction is weaker than the condition imposed in the original Gafni–Bertsekas scheme of [15], namely that

$$N = \hat{C}^*,$$

where

$$\hat{C} = \{w : \forall j \in \hat{\mathcal{J}}, \langle w, a^j \rangle \leq 0\}$$

and

$$\hat{\mathcal{J}} \supset \mathcal{J}(u).$$

In particular, the latter condition does not admit arbitrary closed subspaces $N \supset C(u)^*$.

When the cone N is a subspace, (1) reduces to

$$(3a) \quad u^{i+1} = P_\Omega(u^i + \sigma^i v^i)$$

with

$$\begin{aligned}
(3b) \quad & v = v_N + v_T, \\
(3c) \quad & v_N = -s_N P_N \nabla J(u), \\
(3d) \quad & v_T = -S_T P_T \nabla J(u), \\
(3e) \quad & N = \text{a closed subspace} \supset K(u), \\
(3f) \quad & K(u) = \{w : \forall v \in \Omega, \langle w, v - u \rangle \leq 0\}, \\
(3g) \quad & T = N^\perp, \\
(3h) \quad & \mu_3 \geq s_N \geq \mu_2 > 0, \\
& S_T = \text{a bounded linear map from } T \text{ into } T, \text{ such that} \\
(3i) \quad & \langle P_T \nabla J(u), S_T P_T \nabla J(u) \rangle \geq \mu_0 \|P_T \nabla J(u)\|^2,
\end{aligned}$$

and

$$(3j) \quad \|S_T\| \leq \mu_1,$$

$\sigma = \max s$, subject to

$$\frac{s}{\alpha} \in \{1, \beta, \beta^2, \dots\},$$

and

$$\begin{aligned}
(3k) \quad & J(u) - J(P_\Omega(u + sv)) \geq \delta \{(s_N s)^{-1} \|u + sv_T - P_\Omega(u + sv)\|^2 \\
& - \langle P_T \nabla J(u), v_T \rangle s\}.
\end{aligned}$$

As before, further restrictions on the subspaces N in (3) are dictated by global and local convergence considerations. The subspace decomposition and scaling procedure in [13], [17] applies to closed convex Ω with representations

$$(4a) \quad \Omega = \{u \in \mathcal{U} : g_i(u) \leq 0, j = 1, \dots, m\}.$$

With ϵ_0 and θ fixed in $(0, \infty)$ and $(1, \infty)$ respectively, put

$$\begin{aligned}
(4b) \quad & d(u) = \|u - P_\Omega(u - \nabla J(u))\|, \\
(4c) \quad & \epsilon(u) = \min\{\epsilon_0, d(u)\}, \\
(4d) \quad & \mathcal{J}_0(u) = \{j : g_j(u) = 0\}, \\
(4e) \quad & \mathcal{J}(u) = \{j : g_j(u) \geq -\theta \|\nabla g_j(u)\| \epsilon(u)\} \supset \mathcal{J}_0(u), \\
(4f) \quad & \mathcal{G}_0(u) = \{\nabla g_j(u)\}_{j \in \mathcal{J}_0(u)}, \\
(4g) \quad & \mathcal{G}(u) = \{\nabla g_j(u)\}_{j \in \mathcal{J}(u)} \supset \mathcal{G}_0(u), \\
(4h) \quad & N_0(u) = \text{span}[\mathcal{G}_0(u) \cup \{0\}], \\
(4i) \quad & N(u) = \text{span}[\mathcal{G}(u) \cup \{0\}] \supset N_0(u),
\end{aligned}$$

and require that

$$(4j) \quad N \supset N(u)$$

at each $u \in \Omega$ (note that the sets $\mathcal{J}_o(u)$, $\mathcal{J}(u)$, $\mathcal{G}_o(u)$, and $\mathcal{G}(u)$ may be empty at certain $u \in \Omega$).

The SGP scheme (3), (4) is designed to admit the convergence-accelerating Newtonian scaling operators of [16]. In this construction, we set

$$\begin{aligned} (5a) \quad & N = N(u), \\ (5b) \quad & T = T(u) = N(u)^\perp, \end{aligned}$$

and

$$(5c) \quad S_T = S_T(u) \triangleq E_T L_T^{-1}$$

with

$$\begin{aligned} (5d) \quad & E_T = P_T E|_T \\ (5e) \quad & L_T = P_T L|_T \\ (5f) \quad & E = I + s_N \Lambda \\ (5g) \quad & L = \nabla^2 J(u) + \Lambda \\ (5h) \quad & \Lambda = \sum_{j=1}^m \lambda_j(u) \nabla^2 g_j(u), \end{aligned}$$

where the $\lambda_j(u)$'s are obtained by solving the linear equations

$$\begin{aligned} (5i) \quad & \sum_{j=1}^m \lambda_j(u) \nabla g_j(u) = -P_N \nabla J(u), \\ (5j) \quad & \lambda_j(u) = 0, \quad j \notin \mathcal{J}(u). \end{aligned}$$

The associated full step Newtonian projection (NP) iteration in [16] sets $\sigma = 1$ in place of (3k), and implements (3a)–(3h), (4), and (5). Since the resulting NP sequences $\{u^i\}$ are locally superlinearly convergent to nonsingular minimizers when (4a) satisfies a standard constraint qualification (see below), it seems likely that a similar result can hold for Newtonian SGP processes (3), (4) employing the operators (5). This is shown in §4 when $\alpha = 1$, $\delta \in (0, 1/2)$ and the scaling paramter s_N satisfies the auxiliary rule

$$(6a) \quad s_N = s_N(u) \triangleq \begin{cases} 1, & \text{if } b(u) = 0 \\ \min \left\{ 1, \frac{a(u)}{b(u)} \right\}, & \text{if } b(u) \neq 0, \end{cases}$$

where

$$\begin{aligned} (6b) \quad & a(u) = (1 - 2\delta) \langle P_T \nabla J(u), L_T^{-1} P_T \nabla J(u) \rangle, \\ (6c) \quad & b(u) = \|\Lambda_T L_T^{-1} P_T \nabla J(u)\|^2 + |\langle P_T \nabla J(u), \Lambda_T L_T^{-1} P_T \nabla J(u) \rangle|. \end{aligned}$$

Under these circumstances, the construction (5) is compatible with (3h)–(3j) near nonsingular local minimizers \bar{u} , and the rule (3k) eventually produces unit steps $\sigma^i = 1$ for SGP sequences $\{u^i\}$ converging to \bar{u} , i.e., the Newtonian SGP scheme (3)–(6) eventually reduces to the NP iteration in [16].

In [17], [16] and in the following local convergence analysis, it is assumed that (4a) satisfies the following constraint qualifications.

CONSTRAINT QUALIFICATION Q. *Either*

(i) *the functions g_j in (4a) are affine, i.e.,*

$$g_j(u) = \langle a^j, u \rangle - b^j,$$

with $a^j \in \mathcal{U}$ and $b^j \in \mathbb{R}$, or else

(ii) *(4a) is a normal representation for Ω , i.e.,*

$$\forall u \in \Omega, \mathcal{G}_0(u) \text{ is linearly independent.}$$

In what follows, we will refer to this constraint qualification as Q.

As in [17], a point $\bar{u} \in \Omega$ is said to be *stationary* if and only if

$$-\nabla J(\bar{u}) \in K_\Omega(\bar{u}),$$

or, equivalently,

$$d(\bar{u}) = 0.$$

In convex sets Ω , every local minimizer of J is stationary. A stationary point $\bar{u} \in \Omega$ is said to be *nondegenerate* if and only if

$$-\nabla J(\bar{u}) \in \text{ri}K_\Omega(\bar{u}),$$

where ri denotes the interior of $K_\Omega(u)$ relative to its closed affine hull.

3. Local convergence theorems for general SGP methods. The descent property and fixed point characterization established by Lemma 3 and its corollary in [17] implies that J must be constant on the limit point set for any SGP sequence $\{u^i\}$ generated by (1), and that no local maximizer \bar{u} can be a limit point, except trivially when $u^i = \bar{u}$ for some i . Moreover, while the descent property does not rule out nontrivial subsequential limits at saddles or other spurious stationary points, such occurrences are atypical. These observations, together with the global convergence analysis in [17], and Theorem 1 and Lemma 1 below, suggest that SGP sequences produced by (1), (2) or (3), (4) either have no limit points or are likely to converge to some local minimizer. However, it has not yet been shown that convergence to any *particular* local minimizer \bar{u} must occur if u^i is eventually close enough to \bar{u} . This *asymptotic stability* question is addressed here for (1), (2) and (3), (4) in polyhedral convex Ω satisfying the constraint qualification Q(i), and for (3), (4) in closed convex Ω satisfying Q(ii). The principal results obtained for (1), (2) extend Proposition 3 in [15] in several directions, while the results for (3), (4) are entirely new.

The following development draws on the terminology and theorems in [8], [9].

DEFINITION 1. (i) A fixed point \bar{u} for the SGP iteration (1) is *stable* if and only if for each $\epsilon > 0$, there is a corresponding $\Delta \in (0, \epsilon]$ such that for all sequences $\{u^i\}$ generated by (1),

$$\|u^1 - \bar{u}\| \leq \Delta \Rightarrow \forall i, \quad \|u^i - \bar{u}\| \leq \epsilon.$$

(ii) A fixed point \bar{u} for the SGP iteration (1) is *asymptotically stable* if and only if \bar{u} is stable and there is a $\Delta > 0$ such that for all $\{u^i\}$ generated by (1),

$$\|u^1 - \bar{u}\| \leq \Delta \Rightarrow \lim_{i \rightarrow \infty} u^i = \bar{u}.$$

THEOREM 1. *If \bar{u} is an asymptotically stable fixed point for the SGP iteration (1), then \bar{u} is a proper (i.e., strict) local minimizer and an isolated stationary point for J in Ω .*

Proof. Since every stationary point is a fixed point of (1) [17], \bar{u} cannot be asymptotically stable if stationary points accumulate at \bar{u} . Therefore, suppose that \bar{u} is an isolated stationary point but is not a proper local minimizer. Then every neighborhood of \bar{u} in Ω contains a nonstationary point u^1 at which $J(u^1) \leq J(\bar{u})$. For any sequence $\{u^i\}$ that begins at u^1 and is generated by (1), the descent property insures that for all $i \geq 2$,

$$J(u^i) \leq J(u^2) < J(u^1) \leq J(\bar{u}).$$

Since J is continuous at \bar{u} , it follows that $\{u^i\}$ cannot converge to \bar{u} . Thus if \bar{u} is not an isolated stationary point or \bar{u} is not a proper local minimizer, then \bar{u} is not an asymptotically stable fixed point for (1). \square

Note 1. An example in [18] shows that asymptotically stable proper local minimizers can accumulate at a proper local minimizer \bar{u} ; for certain \bar{u} 's of this type, the sequences generated by (1) cannot converge to \bar{u} from *any* starting point $u^1 \neq \bar{u}$, i.e., \bar{u} cannot be computed with (1).

Our objective now is first to extend the convergence assertion in Proposition 3 of [15], and then to establish several stronger partial converses of Theorem 1 for SGP iterations (1), (2) and (3), (4).

Definition 2. (i) \bar{u} is a *uniformly proper local minimizer* for J in Ω if and only if $\bar{u} \in \Omega$ and there is a $\rho > 0$ and a nondecreasing positive-definite function $a : [0, \rho] \rightarrow [0, \infty]$ such that for all $u \in \Omega$

$$\|u - \bar{u}\| \leq \rho \Rightarrow J(u) - J(\bar{u}) \geq a(\|u - \bar{u}\|).$$

(ii) \bar{u} is a *uniformly isolated stationary point* for J in Ω if and only if $d(\bar{u}) \triangleq \|\bar{u} - P_\Omega(\bar{u} - \nabla J(\bar{u}))\| = 0$, and \bar{u} is a uniformly proper local minimizer of $d(u)$ in Ω . More generally, \bar{u} is a uniformly isolated zero of $f : \mathcal{U} \rightarrow [0, \infty)$ in a closed subset $\mathcal{A} \subset \Omega$ if and only if $\bar{u} \in \mathcal{A}$, $f(\bar{u}) = 0$ and \bar{u} is a uniformly proper local minimizer of f in \mathcal{A} .

Note 2. If $\dim \mathcal{U} < \infty$, then every proper local minimizer of a continuous function J is uniformly proper, every isolated stationary point of a continuously differentiable function J is uniformly isolated, and every isolated zero of a continuous function f is uniformly isolated [8].

LEMMA 1. *Let \bar{u} be a limit point for a sequence $\{u^i\}$ generated by an SGP iteration (1). If \bar{u} is a uniformly proper local minimizer of J in Ω , then $\{u^i\}$ must converge to \bar{u} .*

Proof. Let J , ρ , and $a(\cdot)$ meet the condition in part (i) of Definition 2. Suppose that $\{u^i\}$ does not converge to \bar{u} . Since $u^i - u^{i+1} \rightarrow 0$ (Lemma 4 in [17]), there is an $\epsilon > 0$ and an infinite set Z_1 of positive integers such that for all i

$$\begin{aligned} i \in Z_1 &\Rightarrow u^i \in \{u \in \Omega : \rho \geq \|u - \bar{u}\| \geq \epsilon\} \\ &\Rightarrow J(u^i) - J(\bar{u}) \geq a(\epsilon) > 0. \end{aligned}$$

On the other hand, the descent property insures that

$$\lim_{i \rightarrow \infty} J(u^i) = J(\bar{u}).$$

This contradiction proves that $\{u^i\}$ must converge to \bar{u} . \square

Note 3. The convergence assertion in Proposition 3 of [15] may be seen as a corollary of Lemma 1 (see Note 2).

The strongest local convergence theorems for unconstrained variable metric gradient methods in $\Omega = \mathcal{U}$ have been proved for a special type of uniformly proper local minimizer/uniformly isolated stationary point, namely, the so-called *nonsingular* local minimizer \bar{u} where ∇J vanishes and $\nabla^2 J$ is continuous and coercive; at such points, we have

$$(7a) \quad J(u) - J(\bar{u}) \geq c_1 \|u - \bar{u}\|^2$$

and

$$(7b) \quad \|\nabla J(u)\| \geq c_2 \|u - \bar{u}\|$$

for some $c_1 > 0$, $c_2 > 0$ and all u near \bar{u} [8]. This notion of nonsingularity has a natural extension in Definition 3 below, and once again occupies a central position in the local convergence theories for (1), (2) and (3), (4) in closed convex Ω with representations (4a) satisfying the constraint qualification Q. Observe first that if (4a) satisfies Q, then at each stationary point \bar{u} , there is a Karush–Kuhn–Tucker (KKT) multiplier vector $\bar{\lambda} \in \mathbb{R}^m$ such that

$$(8a) \quad -\nabla J(\bar{u}) = \sum_{j=1}^m \bar{\lambda}_j \nabla g_j(\bar{u}),$$

$$(8b) \quad \bar{\lambda}_j \geq 0, \quad j = 1, \dots, m,$$

$$(8c) \quad (\bar{\lambda}_j > 0 \Rightarrow j \in \mathcal{J}_0(\bar{u})), \quad j = 1, \dots, m$$

with $\mathcal{J}_0(u)$ defined in (4); moreover, it has been shown in [10] that

$$(9) \quad \begin{aligned} &\bar{u} \text{ is a nondegenerate stationary point} \Leftrightarrow \\ &\exists \bar{\lambda} \in \mathbb{R}^m, \bar{\lambda} \text{ satisfies (8) and } (j \in \mathcal{J}_0(\bar{u}) \Rightarrow \bar{\lambda}_j > 0), \quad j = 1, \dots, m. \end{aligned}$$

In particular, if Q(ii) holds, then $\bar{\lambda}$ is unique. At $u \in \Omega$, put

$$T_0(u) = N_0(u)^\perp$$

with $N_0(u)$ defined in (4). When J and the functions g_j in (4a) are twice Fréchet differentiable at u , put

$$(10) \quad L(u, \lambda) = \begin{cases} \nabla^2 J(u), & \text{if Q(i) holds} \\ \nabla^2 J(u) + \sum_{j=1}^m \lambda_j \nabla^2 g_j(u), & \text{if Q(ii) holds.} \end{cases}$$

DEFINITION 3. Assume that Ω is a closed convex set with a representation (4a) that satisfies the constraint qualification Q. A local minimizer (respectively, stationary point) \bar{u} for J in Ω is *nonsingular* if and only if \bar{u} is nondegenerate, J and the functions g_j in (4a) are twice continuously differentiable at \bar{u} , and the reduced Lagrangian Hessian,

$$P_{T_0(\bar{u})} L(\bar{u}, \bar{\lambda})|_{T_0(\bar{u})},$$

is coercive (respectively, bijective) for any KKT multiplier $\bar{\lambda}$ satisfying the right-hand side of (9). (Note that either $\bar{\lambda}$ is unique or $L(\bar{u}, \bar{\lambda})$ does not depend on the choice of $\bar{\lambda}$. Therefore, in all cases, $L(\bar{u}, \bar{\lambda})$ is uniquely prescribed at \bar{u} .)

LEMMA 2. *Suppose that \bar{u} is a nonsingular local minimizer for J in a closed convex set Ω with a representation (4a) that satisfies the constraint qualification Q, and let*

$$\mathcal{A}(\bar{u}) = \{u \in \Omega : \mathcal{J}_0(u) = \mathcal{J}_0(\bar{u})\}.$$

Then for some $\rho > 0$, $c_1 > 0$, and $c_2 > 0$, and all u ,

$$(11a) \quad u \in B(\bar{u}, \rho) \cap \Omega \Rightarrow J(u) - J(\bar{u}) \geq c_1 \|u - \bar{u}\|^2,$$

$$(11b) \quad u \in B(\bar{u}, \rho) \cap \mathcal{A}(\bar{u}) \Rightarrow \|P_{T_0(u)} \nabla J(u)\| \geq c_2 \|u - \bar{u}\|.$$

Proof. The estimate (11a) can be deduced with (9) and the general Banach space sufficiency proof technique in [19, pp. 191–192]. The estimate (11b) is established in [9] when Q(ii) holds. Suppose instead that (4a) satisfies Q(i). Observe that for all $u \in \mathcal{A}(\bar{u})$, one has $\mathcal{J}_0(u) = \mathcal{J}_0(\bar{u})$ and consequently $T_0(u) = T_0(\bar{u})$. Since $P_{T_0(\bar{u})} \nabla J(u)$ and $P_{T_0(\bar{u})} \nabla^2 J(u)|_{T_0(\bar{u})}$ can be seen as the gradient and Hessian of a translate of J restricted to the Hilbert space $T_0(\bar{u})$, assertion (11b) now follows from (7b) and the coercivity of $P_{T_0(\bar{u})} \nabla^2 J(\bar{u})|_{T_0(\bar{u})} = P_{T_0(\bar{u})} L(\bar{u}, \bar{\lambda})|_{T_0(\bar{u})}$.

Note 4. In view of (11), every nonsingular local minimizer is automatically a uniformly proper local minimizer and a uniformly isolated zero of the function $\|P_{T_0(u)} \nabla J(u)\|$ in $\mathcal{A}(\bar{u})$.

One additional property of nondegenerate stationary points is needed to establish a link with hypothesis (B) in the local convergence analysis of [15], and to develop an analogous local theory for (3), (4) in nonpolyhedral sets.

LEMMA 3. *Let \bar{u} be a nondegenerate stationary point for J in a closed convex set Ω with a representation (4a) satisfying the constraint qualification Q. Assume that the functions g_j in (4a) are continuously Fréchet differentiable at \bar{u} . Let α_1 and α_2 be fixed positive numbers, with $\alpha_1 < \alpha_2$. Then for some corresponding $\epsilon > 0$, and for all h and s ,*

$$(12) \quad \|h\| \leq \epsilon \quad \text{and} \quad s \in [\alpha_1, \alpha_2] \Rightarrow P_\Omega(\bar{u} - s \nabla J(\bar{u}) + sh) \in \mathcal{A}(\bar{u}).$$

Proof. Put $\bar{v} = -\nabla J(\bar{u}) \in \text{ri } K(\bar{u})$. Suppose that Q(i) holds. Then for some $\epsilon > 0$, and all h and s ,

$$\begin{aligned} \|h\| \leq \epsilon \quad \text{and} \quad s \in [\alpha_1, \alpha_2] &\Rightarrow (\bar{u} + s P_{T_0(\bar{u})} h) \in \mathcal{A}(\bar{u}) \quad \text{and} \quad s(\bar{v} + P_{N_0(\bar{u})} h) \in K(\bar{u}) \\ &\Rightarrow (\bar{u} + s P_{T_0(\bar{u})} h) \in \mathcal{A}(\bar{u}) \quad \text{and} \quad s(\bar{v} + P_{N_0(\bar{u})} h) \in K(\bar{u} + s P_{T_0(\bar{u})} h) \\ &\Rightarrow P_\Omega(\bar{u} + s \bar{v} + sh) = (\bar{u} + s P_{T_0(\bar{u})} h) \in \mathcal{A}(\bar{u}). \end{aligned}$$

On the other hand, if Q(ii) holds, then an argument similar to that used in the proof of Lemma 1 in [9] will show that $\mathcal{J}_0(P_\Omega(\bar{u} + s \bar{v} + sh)) = \mathcal{J}_0(\bar{u})$ for $s \in [\alpha_1, \alpha_2]$ and $\|h\|$ sufficiently small. Thus, note that since the g_j 's are continuous at \bar{u} , and since

$$\|P_\Omega(\bar{u} + s \bar{v} + sh) - \bar{u}\| = \|P_\Omega(\bar{u} + s \bar{v} + sh) - P_\Omega(\bar{u} + s \bar{v})\| \leq s \|h\|,$$

it follows that

$$\exists \epsilon > 0, \quad \forall s \in [\alpha_1, \alpha_2], \quad \forall h \in B(0, \epsilon_1), \quad \mathcal{J}_0(P_\Omega(\bar{u} + s \bar{v} + sh)) \subset \mathcal{J}_0(\bar{u}).$$

To prove the reverse inclusion, note that if $\mathcal{J}_0(\bar{u}) = \emptyset$ then trivially $\mathcal{J}_0(\bar{u}) \subset \mathcal{J}_0(P_\Omega(\bar{u} + s\bar{v} + sh))$ for all s, h . Therefore, suppose that $\mathcal{J}_0(\bar{u}) \neq \emptyset$. With reference to (9) there is a $\bar{\lambda} \in \mathbb{R}^m$ such that

$$(13a) \quad (j \notin \mathcal{J}_0(\bar{u}) \Rightarrow \bar{\lambda}_j = 0), \quad j = 1, \dots, m,$$

$$(13b) \quad b \triangleq \min_{j \in \mathcal{J}_0(\bar{u})} \bar{\lambda}_j > 0,$$

and

$$(13c) \quad \bar{v} = \sum_{j=1}^m \bar{\lambda}_j \nabla g_j(\bar{u}).$$

Furthermore, Q(ii) insures that for all s, h , there is a corresponding $\ell(s, h) \in \mathbb{R}^m$ such that

$$(14a) \quad \ell_j(s, h) \geq 0, \quad j = 1, \dots, m,$$

$$(14b) \quad (j \notin \mathcal{J}_0(P_\Omega(\bar{u} + s\bar{v} + sh)) \Rightarrow \ell_j(s, h) = 0), \quad j = 1, \dots, m,$$

and

$$(14c) \quad \bar{u} + s\bar{v} + sh - P_\Omega(\bar{u} + s\bar{v} + sh) = \sum_{j=1}^m \ell_j(s, h) \nabla g_j(P_\Omega(\bar{u} + s\bar{v} + sh));$$

therefore,

$$(14d) \quad \begin{aligned} & 2s\|h\| + \|\ell\|_2 \|\nabla g(P_\Omega(\bar{u} + s\bar{v} + sh)) - \nabla g(\bar{u})\|_2 \\ & \geq \left\| \sum_{j=1}^m (\ell_j(s, h) - s\bar{\lambda}_j) \nabla g_j(\bar{u}) \right\| \end{aligned}$$

with

$$\|\nabla g(P_\Omega(\bar{u} + s\bar{v} + sh)) - \nabla g(\bar{u})\|_2 \triangleq \left(\sum_{j=1}^m \|\nabla g_j(P_\Omega(\bar{u} + s\bar{v} + sh)) - \nabla g_j(\bar{u})\|^2 \right)^{1/2}.$$

For $s \in [\alpha_1, \alpha_2]$ and $\|h\|$ sufficiently small, it will now be shown that $\ell_j(s, h)$ must be positive for all $j \in \mathcal{J}_0(\bar{u})$, and hence that $\mathcal{J}_0(\bar{u}) \subset \mathcal{J}_0(P_\Omega(\bar{u} + s\bar{v} + sh))$. Observe first that Q(ii) and the continuity of the gradients $\nabla g_j(\cdot)$ at \bar{u} imply that

$$\exists \epsilon_2 \in [0, \epsilon_1], \quad \exists \gamma > 0, \quad \forall s \in [\alpha_1, \alpha_2], \quad \forall h \in B(0, \epsilon_2), \quad \forall c \in \mathbb{R}^m,$$

$$\gamma \left\| \sum_{j \in \mathcal{J}_0(\bar{u})} c_j \nabla g_j(P_\Omega(\bar{u} + s\bar{v} + sh)) \right\| \geq \left(\sum_{j \in \mathcal{J}_0(\bar{u})} c_j^2 \right)^{1/2}.$$

Therefore, in view of (14), there is an $\epsilon \in (0, \epsilon_2]$ such that for all s, h ,

$$\begin{aligned} \|h\| \leq \epsilon \quad \text{and} \quad s \in [\alpha_1, \alpha_2] & \Rightarrow \frac{1}{2} \gamma^{-1} s b \geq \gamma^{-1} \max_{j \in \mathcal{J}_0(\bar{u})} |\ell_j(s, h) - s\bar{\lambda}_j| \\ & \Rightarrow \forall j \in \mathcal{J}_0(\bar{u}), \quad \ell_j(s, h) \geq s \left(\bar{\lambda}_j - \frac{b}{2} \right) > 0 \\ & \Rightarrow \mathcal{J}_0(\bar{u}) \subset \mathcal{J}_0(P_\Omega(\bar{u} + s\bar{v} + sh)). \quad \square \end{aligned}$$

Note 5. Suppose that Ω is a polyhedral convex set with a representation (4a) satisfying Q(i) and let (12) hold with $\alpha_1 = \alpha_2 = 1$. Then for all h ,

$$\begin{aligned} h \in N_0(\bar{u}) \quad \text{and} \quad \|h\| \leq \epsilon &\Rightarrow P_\Omega(\bar{u} - \nabla J(\bar{u}) + h) \in \mathcal{A}(\bar{u}) \\ &\Rightarrow \exists \xi \in \mathcal{A}(\bar{u}), \quad (\bar{u} - \nabla J(\bar{u}) + h - \xi) \in K(\xi) \\ &\Rightarrow \exists \xi \in \mathcal{A}(\bar{u}), \quad [(\bar{u} - \xi) + (-\nabla J(\bar{u}) + h)] \in K(\bar{u}). \end{aligned}$$

Since $\bar{u} - \xi \in T_0(\bar{u}) = N_0(\bar{u})^\perp$ and $-\nabla J(\bar{u}) + h \in N_0(\bar{u}) \supset K(\bar{u})$, it now follows that for all h

$$h \in N_0(\bar{u}) \quad \text{and} \quad \|h\| \leq \epsilon \Rightarrow -\nabla J(\bar{u}) + h \in K(\bar{u}).$$

Thus, when Q(i) holds, the converse of Lemma 3 is also true, and the nondegeneracy condition $-\nabla J(\bar{u}) \in \text{ri } K(\bar{u})$ is *equivalent* to hypothesis (B) in [15]. This can also be shown by applying Theorem 2.8 in [10], after noting that $\mathcal{A}(\bar{u})$ is the relative interior of a quasipolyhedral face when Q(i) is satisfied. Similarly, when Q(ii) holds, it turns out that $\mathcal{A}(\bar{u})$ is an *open facet* \mathcal{F} in Ω [8], and for any such \mathcal{F} one can prove that

$$\text{int } P_\Omega^{-1}[\mathcal{F}] = \mathcal{F} + \text{ri } K(\mathcal{F}),$$

where $K(\mathcal{F}) = K(u)$ for all $u \in \mathcal{F}$; once again, Lemma 3 and its converse are corollaries of this more general result.

The following lemma leads immediately to a stability condition for fixed points of the SGP iterations (1), (2) and (3), (4).

LEMMA 4. *Let \bar{u} be a nondegenerate stationary point in a closed convex set Ω with a representation (4a) that satisfies the constraint qualification Q. Suppose that J and the functions g_j in (4a) are continuously Fréchet differentiable at \bar{u} . Then*

$$(15) \quad \exists \rho > 0, \quad \forall u \in B(\bar{u}, \rho) \cap \Omega, \quad [P_\Omega(u - \nabla J(u)) \in \mathcal{A}(\bar{u}) \quad \text{and} \quad \mathcal{J}(u) = \mathcal{J}(\bar{u})]$$

with $\mathcal{J}(u)$ defined by (4). Furthermore, if Q(i) is satisfied then (15) also holds with $\mathcal{J}(u)$ defined by (2).

Proof. Suppose that Q(ii) holds. Then, by Lemma 3 and the continuity of $\nabla J(\cdot)$ at \bar{u} ,

$$(16) \quad \exists \rho_1 > 0, \quad \forall u \in B(\bar{u}, \rho_1), \quad P_\Omega(u - \nabla J(u)) \in \mathcal{A}(\bar{u}).$$

The remaining portion of (15) is established by Lemma 3.5 in [16], whose proof bears repeating here. Observe first that $d(\cdot)$, $g_j(\cdot)$, and $\nabla g_j(\cdot)$ are continuous at \bar{u} , with $d(\bar{u}) = 0$. Hence for any fixed θ there is a $\rho_2 \in (0, \rho_1]$ and a $c < 0$ such that for all u ,

$$(17) \quad \begin{aligned} u \in B(\bar{u}, \rho_1) \cap \Omega &\Rightarrow \forall j \notin \mathcal{J}_0(\bar{u}), \quad g_j(u) \leq \frac{c}{2} \leq -\theta \|\nabla g_j(u)\| \epsilon(u) \\ &\Rightarrow \mathcal{J}(u) \subset \mathcal{J}_0(\bar{u}) \end{aligned}$$

with $\mathcal{J}(u)$ defined by (4). To prove the reverse inclusion, note that by (16) and the mean value theorem, we have, for all u ,

$$(18) \quad \begin{aligned} u \in B(\bar{u}, \rho_2) \cap \Omega &\Rightarrow \forall j \in \mathcal{J}_0(\bar{u}), \quad g_j(P_\Omega(u - \nabla J(u))) = 0 \\ &\Rightarrow \forall j \in \mathcal{J}_0(\bar{u}), \quad \exists \xi^j \in \Omega, \\ g_j(u) &\geq -(\|\nabla g_j(u)\| + \|\nabla g_j(\xi^j) - \nabla g_j(u)\|)d(u), \quad \text{and} \quad \|\xi^j - u\| \leq d(u). \end{aligned}$$

Furthermore, by Q(ii), $\|\nabla g_j(\bar{u})\| > 0$ for all $j \in \mathcal{J}_0(\bar{u})$. Consequently for any fixed $\theta > 1$, there is a $\rho \in (0, \rho_2]$ such that for all u ,

$$(19) \quad \begin{aligned} u \in B(\bar{u}, \rho) \cap \Omega &\Rightarrow \forall j \in \mathcal{J}_0(\bar{u}), \quad g_j(u) \geq -\theta \|\nabla g_j(u)\| \epsilon(u) \\ &\Rightarrow \mathcal{J}_0(\bar{u}) \subset \mathcal{J}(u) \end{aligned}$$

with $\mathcal{J}(u)$ defined by (4).

Now suppose that Q(i) holds. In this case, (16) follows from Lemma 3 as before, and (17) is therefore true once again for any fixed θ , and thus for $\mathcal{J}(u)$ defined by (2) or (4). Conversely, since affine g_j 's have constant gradients, $\nabla g_j = a^j$, the estimates in (16) establish (17) for any fixed $\theta \geq 1$ and hence for $\mathcal{J}(u)$ defined by (2) or (4). \square

Note 6. If τ is fixed in $(0, \infty)$ and $\epsilon(\cdot)$ is replaced in (2) and (4) by any continuous measure of nonstationarity satisfying

$$\epsilon(u) \geq \|u - P_\Omega(u - \tau \nabla J(u))\|$$

near nondegenerate stationary points \bar{u} , then the estimates in (15) continue to hold provided $P_\Omega(u - \nabla J(u))$ is replaced by $P_\Omega(u - \tau \nabla J(u))$. Since this alteration has no impact on the subsequent local convergence analyses, the function $\epsilon(\cdot)$ in (2) and (4) is merely the prototype for a large class of admissible nonstationarity measures for SGP iterations (1), (2) and (3), (4). In fact, we wonder whether the restriction $\theta > 1$ imposed in (4) is again superfluous, as it was in the global analyses of [17]. This would certainly be true if there is a $\theta > 1$ and sufficiently small positive numbers τ and ρ such that for all u ,

$$u \in B(\bar{u}, \rho) \cap \Omega \Rightarrow \theta \|u - P_\Omega(u - \tau \nabla J(u))\| \leq \|u - P_\Omega(u - \nabla J(u))\|.$$

Unfortunately, this proposition is generically false for nondegenerate stationary points in the relative boundary of sets (4a) satisfying Q. For example, in $\mathcal{U} = \mathbb{R}^2$ let Ω be the nonpositive orthant $\{(u_1, u_2) : u_1 \leq 0, u_2 \leq 0\}$ and let $J(u) = -u_1 - u_2$. Then Ω satisfies Q, and J has a nondegenerate stationary point (and global minimizer) at $\bar{u} = (0, 0)$. At boundary points $(u_1, 0)$ near \bar{u} in Ω , it can be seen that for all τ ,

$$\begin{aligned} \sqrt{|u_1|} &\leq \tau \leq 1 \Rightarrow P_\Omega(u - \tau \nabla J(u)) = \bar{u} \\ &\Rightarrow \|u - P_\Omega(u - \tau \nabla J(u))\| = \|u - P_\Omega(u - \nabla J(u))\|. \end{aligned}$$

Thus for any fixed $\theta > 1$ and $\tau \in (0, 1)$, every deleted neighborhood of \bar{u} in Ω contains points at which

$$\theta \|u - P_\Omega(u - \theta \nabla J(u))\| > \|u - P_\Omega(u - \nabla J(u))\|.$$

Since similar counterexamples are readily constructed in nonpolyhedral sets (4a), there seems to be no easy way around the restriction $\theta > 1$ in (4).

LEMMA 5. *Assume that Ω is a closed convex set with a representation (4a) satisfying the constraint qualification Q. Let \bar{u} be a uniformly proper local minimizer and a nondegenerate stationary point for J in Ω , and suppose that J and the functions g_j in (4a) are continuously Fréchet differentiable at \bar{u} . Then \bar{u} is a stable fixed point for the SGP iteration (3), (4). Furthermore, if Q(i) is satisfied then \bar{u} is also a stable fixed point for (1), (2).*

Proof. By Lemma 3 and its corollary in [17], \bar{u} is a fixed point of the SGP iterations in question, and stability will follow from the descent property and a Lyapunov

construction, provided the associated maps $u \rightarrow P_\Omega(u + \sigma v)$ are continuous at \bar{u} , uniformly in the parameters N, s_N and S_T admitted by (3), (4) or (1), (2), respectively.

For any map $u \rightarrow P_\Omega(u + \sigma v)$ satisfying (3), (4), we have

$$\begin{aligned}
 (20a) \quad \|P_\Omega(u + \sigma v) - \bar{u}\| &= \|P_\Omega(u + \sigma v) - P_\Omega(\bar{u} - \sigma s_N \nabla J(\bar{u}))\| \\
 &\leq \|u - \bar{u} + \sigma(v + s_N \nabla J(\bar{u}))\| \\
 &\leq \|u - \bar{u}\| + \alpha(\mu_1 + \mu_3)\|P_{T(u)}\nabla J(u)\| + \alpha\mu_3\|\nabla J(u) - \nabla J(\bar{u})\|
 \end{aligned}$$

since $N^\perp = T \subset T(u)$. If Q(ii) holds, then Lemma 3.2 and Theorem 3.1 in [15] establish that

$$(20b) \quad \lim_{\substack{u \rightarrow \bar{u} \\ u \in \Omega}} P_{T(u)}\nabla J(u) = 0.$$

Hence $u \rightarrow P_\Omega(u + \sigma v)$ is seen to be continuous at \bar{u} , uniformly in N, s_N and S_T when Q(ii) holds. Given any $\epsilon > 0$, it is now possible to construct a corresponding set $\mathcal{I}_\epsilon \subset B(\bar{u}, \epsilon) \cap \Omega$ such that \mathcal{I}_ϵ is invariant under every such map, and \bar{u} lies in the interior of \mathcal{I}_ϵ relative to Ω ; evidently, this construction will prove that \bar{u} is stable for (3), (4). With reference to Definition 2, choose $\rho > 0$ and $a(\cdot)$ so that for all u ,

$$u \in B(\bar{u}, \rho) \cap \Omega \Rightarrow J(u) - J(\bar{u}) \geq a(\|u - \bar{u}\|).$$

Fix ϵ in $(0, \rho]$. Then for some $r > 0$, and all u, N, T, s_N , and S_T ,

$$u \in B(\bar{u}, r) \cap \Omega \quad \text{and} \quad (3), (4) \Rightarrow \|P_\Omega(u + \sigma v) - \bar{u}\| \leq \epsilon.$$

Construct the corresponding set

$$\mathcal{I}_\epsilon = \{u \in \Omega : \|u - \bar{u}\| \leq \epsilon \text{ and } J(u) - J(\bar{u}) < a(r)\}.$$

By the descent property and the properties of $a(\cdot)$, it follows that for all u, N, T, s_N, S_T ,

$$\begin{aligned}
 u \in \mathcal{I}_\epsilon \text{ and } (3), (4) &\Rightarrow (a(\|u - \bar{u}\|) \leq J(u) - J(\bar{u}) < a(r)) \\
 &\quad \text{and } J(P_\Omega(u + \sigma v)) - J(\bar{u}) < a(r) \\
 &\Rightarrow (\|u - \bar{u}\| \leq r, \|P_\Omega(u + \sigma v) - \bar{u}\| \leq \epsilon, \\
 &\quad \text{and } J(P_\Omega(u + \sigma v)) - J(\bar{u}) < a(r)) \\
 &\Rightarrow P_\Omega(u + \sigma v) \in \mathcal{I}_\epsilon.
 \end{aligned}$$

Thus \mathcal{I}_ϵ is invariant under any map $u \rightarrow P_\Omega(u + \sigma v)$ satisfying (3), (4); moreover, since J is continuous at \bar{u} , there is a $\Delta \in (0, r]$ such that

$$B(\bar{u}, \Delta) \cap \Omega \subset \mathcal{I}_\epsilon \subset B(\bar{u}, \epsilon) \cap \Omega.$$

By construction, if $u^1 \in B(\bar{u}, \Delta) \cap \Omega$ and if $\{u^i\}$ is generated by (3), (4), then u^i must remain in \mathcal{I}_ϵ for $i \geq 1$.

Similarly, for any map $u \rightarrow P_\Omega(u + \sigma v)$ satisfying (1), (2), it can be seen that

$$\begin{aligned}
 (21a) \quad \|P_\Omega(u + \sigma v) - \bar{u}\| &\leq \|u - \bar{u}\| + \alpha(\mu_1 + \mu_3)\|P_{C(u)}(-\nabla J(u))\| + \alpha\mu_3\|\nabla J(u) - \nabla J(\bar{u})\|
 \end{aligned}$$

Furthermore, if Q(i) holds, then Lemma 4 implies that for u near \bar{u} in Ω ,

$$(21b) \quad C(u)^* = C(\bar{u})^* = K(\bar{u});$$

therefore,

$$(21c) \quad \begin{aligned} \|P_{C(u)}(-\nabla J(u))\| &= \|P_{C(\bar{u})}(-\nabla J(\bar{u})) + P_{C(\bar{u})}(-\nabla J(u)) - P_{C(\bar{u})}(-\nabla J(\bar{u}))\| \\ &\leq \|\nabla J(u) - \nabla J(\bar{u})\|. \end{aligned}$$

Hence, $u \rightarrow P_\Omega(u + \sigma v)$ is continuous at \bar{u} uniformly in N, s_N , and S_T , and a repetition of the foregoing argument proves that \bar{u} is stable for (1), (2), and hence for (3), (4) (see Note 1 in [17]). \square

Lemma 5 immediately yields a complete characterization of nondegenerate asymptotically stable fixed points for (3), (4) and (1), (2) in finite-dimensional spaces.

THEOREM 2. *Assume that $\dim \mathcal{U} < \infty$ and that Ω is a closed convex set with a representation (4a) satisfying the constraint qualification Q. Let \bar{u} be a nondegenerate stationary point for J in Ω and suppose that J and the functions g_j in (4a) are continuously Fréchet differentiable at \bar{u} . Then \bar{u} is an asymptotically stable fixed point for the SGP iteration (3), (4) if and only if \bar{u} is a proper local minimizer and an isolated stationary point for J in Ω . Furthermore, if Q(i) is satisfied, then the same conclusion holds for (1), (2).*

Proof. By Theorem 1, an asymptotically stable fixed point for any SGP iteration (1) must be a proper local minimizer and an isolated stationary point. Conversely, suppose that \bar{u} is a proper local minimizer. Since J is differentiable at \bar{u} and hence continuous near \bar{u} , it can be seen that \bar{u} is a uniformly proper local minimizer (see Note 2). Since \bar{u} is nondegenerate, Lemma 4 asserts that \bar{u} is stable for (3), (4). Now suppose that for some $\epsilon > 0$, \bar{u} is the only stationary point in $B(\bar{u}, \epsilon) \cap \Omega$. Choose $\Delta \in (0, \epsilon]$ so that every sequence that begins in $B(\bar{u}, \Delta) \cap \Omega$ and is generated by (3), (4) will remain in $B(\bar{u}, \epsilon) \cap \Omega$. All such sequences must converge to \bar{u} , in view of Theorem 2 in [17] and the compactness of $B(\bar{u}, \epsilon) \cap \Omega$. When Q(i) holds, the same argument applies to (1), (2). \square

Two variants of Theorem 2 will now be established in arbitrary real Hilbert spaces \mathcal{U} , with the aid of the following supplement to Lemmas 5–8 in [17].

LEMMA 6. *Let \bar{u} be a nondegenerate stationary point for J in a closed convex set Ω with a representation (4a) that satisfies the constraint qualification Q. Assume that J and the functions g_j in (4a) are continuously Fréchet differentiable at \bar{u} . If Q(ii) holds and v is determined by (3b)–(3j) and (4), then for some $\Delta > 0$, $\rho > 0$ and $\kappa > 0$, and all $\sigma \in (0, \Delta]$ and $u \in B(\bar{u}, \rho) \cap \Omega$,*

$$(22a) \quad \langle x, P_\Omega(u + \sigma v) - u \rangle \geq \langle x, v \rangle \sigma + \|x\| \eta(\sigma, u, v)$$

and

$$(22b) \quad \begin{aligned} \langle \nabla J(u), u - P_\Omega(u + \sigma v) \rangle &\geq [s_N^{-1} \sigma^{-2} \|u + \sigma v_T - P_\Omega(u + \sigma v)\|^2 - \langle P_T \nabla J(u), v_T \rangle] \sigma \\ &\quad + s_N^{-1} \|x\| \eta(\sigma, u, v) \end{aligned}$$

with

$$(22c) \quad x = (v_T - s_N P_T \nabla J(u)) \in T \subset T(u)$$

and

$$(22d) \quad \eta(\sigma, u, v) = -\kappa \|\nabla g(P_\Omega(u + \sigma v)) - \nabla g(u)\|_2 \sigma.$$

Furthermore if Q(i) holds and v is determined by (1b)–(1l) and (2), then for some $\Delta > 0$ and $\rho > 0$, and all $\sigma \in (0, \Delta]$ and $u \in B(\bar{u}, \rho) \cap \Omega$,

$$(23a) \quad \langle x, P_\Omega(u + \sigma v) - u \rangle \geq \langle x, v \rangle \sigma$$

and

$$(23b) \quad \langle \nabla J(u), u - P_\Omega(u + \sigma v) \rangle \geq (s_N \sigma)^{-1} \|u + \sigma v_T - P_\Omega(u + \sigma v)\|^2 + \langle P_{N^*}(-\nabla J(u)), v_T \rangle \sigma$$

with

$$(23c) \quad x = v_T + s_N P_{N^*}(-\nabla J(u)) \in T \subset C(u).$$

Proof. The estimates (22b) and (23b) derive from the inequality (8) in [17] and (22a) and (23a), as explained in the proof of Lemma 2 in [17]. The inequalities (22a) and (23a) are established as follows. Recall first that for all $\sigma \geq 0$, $u \in \Omega$, $v \in \mathcal{U}$,

$$(24a) \quad (u + \sigma v - P_\Omega(u + \sigma v)) \in K(P_\Omega(u + \sigma v))$$

and

$$(24b) \quad \|P_\Omega(u + \sigma v) - \bar{u}\| \leq \|u - \bar{u}\| + \|v\| \sigma$$

Now suppose that Q(ii) holds. An examination of the first part of the proof of Lemma 7 in [17] will show that for some $M > 0$, $\Delta_1 > 0$, $\rho_1 > 0$, $\kappa > 0$, and for all σ, u, v

$$(25) \quad u \in B(\bar{u}, \rho_1) \cap \Omega, \text{ (3b)–(3j), and (4)} \Rightarrow \|v\| \leq (\mu_1 + \mu_3) \|\nabla J(u)\| \leq M$$

and

$$\sigma \in (0, \Delta_1], \quad u \in B(\bar{u}, \rho_1) \cap \Omega, \text{ (3b)–(3j), and (4)} \Rightarrow$$

$$\forall x \in T(u), \langle x, P_\Omega(u + \sigma v) - u \rangle \geq \langle x, v \rangle \sigma + \|x\| \eta(\sigma, u, v) - \left| \langle x, \sum_{j=1}^m c_j \nabla g_j(u) \rangle \right|,$$

where c is a nonnegative vector in \mathbb{R}^m , depending on σ, u, v and satisfying

$$(j \notin \mathcal{J}(P_\Omega(u + \sigma v))) \Rightarrow c_j = 0, \quad j = 1, \dots, m.$$

This much is true for any $\bar{u} \in \Omega$. Moreover, if \bar{u} is a nondegenerate stationary point, then in view of (24b), (25), and Lemma 4, there are numbers $\Delta \in (0, \Delta_1]$ and $\rho \in (0, \rho_1]$, such that for all σ, u, v ,

$$\sigma \in (0, \Delta], \quad u \in B(\bar{u}, \rho) \cap \Omega, \text{ (3b)–(3j) and (4)} \Rightarrow \mathcal{J}(u) = \mathcal{J}_0(\bar{u}) = \mathcal{J}(P_\Omega(u + \sigma v))$$

$$\Rightarrow \sum_{j=1}^m c_j \nabla g_j(u) \in N(u)$$

$$\Rightarrow \forall x \in T(u), \langle x, P_\Omega(u + \sigma v) - u \rangle \geq \langle x, v \rangle \sigma + \|x\| \eta(\sigma, u, v)$$

On the other hand, suppose that Q(i) holds. Then according to Lemma 4 and (24), there are numbers $M > 0$, $\Delta > 0$ and $\rho > 0$ such that for all σ, u, v ,

$$u \in B(\bar{u}, \rho) \cap \Omega, (1B) - (1L), \text{ and } (2) \Rightarrow \|v\| \leq (\mu_1 + \mu_3) \|\nabla J(u)\| \leq M$$

and

$$\begin{aligned} \sigma \in (0, \Delta], u \in B(\bar{u}, \rho) \cap \Omega, (1B) - (1L) \text{ and } (2) &\Rightarrow C(u)^* = K(u) = C(P_\Omega(u + \sigma v))^* \\ &\Rightarrow (u + \sigma v - P_\Omega(u + \sigma v)) \in C(u)^* \\ &\Rightarrow \forall x \in C(u), \langle x, u + \sigma v - P_\Omega(u + \sigma v) \rangle \leq 0. \quad \square \end{aligned}$$

With Lemma 6, it is now possible to prove that the step lengths σ generated by (3), (4) and (1), (2) are bounded away from zero near nondegenerate stationary points \bar{u} , provided J and the constraint functions g_j have locally Lipschitz continuous gradients near \bar{u} . A bound of this sort is needed for the active constraint identification result in Proposition 3 of [15], and plays a similar part here in the proofs of Theorems 3 and 4 to follow.

LEMMA 7. *Let \bar{u} be a nondegenerate stationary point for J in a closed convex set Ω with a representation (4a) satisfying the constraint qualification Q. If Q(ii) holds and the gradients of J and the constraint functions g_j in (4a) are Lipschitz continuous near \bar{u} , then the step lengths σ generated by (3), (4) are bounded away from 0 near \bar{u} . Similarly, if Q(i) holds and the gradient of J is Lipschitz continuous near \bar{u} then the step lengths σ generated by (1), (2) are likewise bounded away from 0.*

Proof. Suppose that Q(ii) holds and σ is determined by (3), (4) at $u \in \Omega$. If u is stationary, then $\sigma = \alpha$. Assume that u is not stationary and that $\sigma \leq \alpha\beta$. Then (3k) implies that

$$(26) \quad \begin{aligned} J(u) - J(P_\Omega(u + \beta^{-1}\sigma v)) &< \delta\{(s_N\beta^{-1}\sigma)^{-1}\|u + \beta^{-1}\sigma v_T - P_\Omega(u + \beta^{-1}\sigma v)\|^2 \\ &\quad - \langle P_T \nabla J(u), v_T \rangle \beta^{-1}\sigma\} \end{aligned}$$

With reference to (10), observe that

$$(27a) \quad \begin{aligned} \|P_\Omega(u + \beta^{-1}\sigma v) - \bar{u}\| &\leq \|u - \bar{u}\| + \beta^{-1}\alpha(\mu_1 + \mu_3)\|P_{T(u)} \nabla J(u)\| \\ &\quad + \beta^{-1}\alpha\|\nabla J(u) - \nabla J(\bar{u})\| \end{aligned}$$

with

$$(27b) \quad \lim_{u \rightarrow \bar{u}} P_{T(u)} \nabla J(u) = 0.$$

Hence, there are numbers $\Delta > 0$, $\rho > 0$, $\kappa > 0$, and $\rho_1 \in (0, \rho]$ such that the estimates (20) hold, and for all $u \in B(\bar{u}, \rho_1) \cap \Omega$,

$$(28a) \quad P_\Omega(u + \beta^{-1}\sigma v) \in B(\bar{u}, \rho) \cap \Omega,$$

$$(28b) \quad J(u) - J(P_\Omega(u + \beta^{-1}\sigma v)) \geq \langle \nabla J(u), u - P_\Omega(u + \beta^{-1}\sigma v) \rangle - \frac{1}{2} L_J \|P_\Omega(u + \beta^{-1}\sigma v) - u\|^2,$$

and

$$(28c) \quad \|\nabla g(P_\Omega(u + \beta^{-1}\sigma v)) - \nabla g(u)\|_2 \leq L_g \|P_\Omega(u + \beta^{-1}\sigma v) - u\|,$$

where L_J and L_g are local Lipschitz constants for ∇J and ∇g . According to (22) and (28), it follows that for all $u \in B(\bar{u}, \rho_1) \cap \Omega$, either $\sigma > \beta \min\{\alpha, \Delta\}$, or

$$\begin{aligned}
 (29a) \quad & (1 + \beta^{-1}\alpha\kappa L_g)\|x\| \|P_\Omega(u + \beta^{-1}\sigma v) - u\| \geq (\|v_T\|^2 + \mu_0\mu_2\|P_T\nabla J(u)\|^2)\beta^{-1}\sigma \\
 & \geq \min\left\{1, \frac{\mu_0\mu_2}{2\mu_3(\mu_1 + \mu_3)}\right\} \|x\|^2\beta^{-1}\sigma \\
 & = \frac{\mu_0\mu_2}{2\mu_3(\mu_1 + \mu_3)} \|x\|^2\beta^{-1}\sigma
 \end{aligned}$$

and

$$\begin{aligned}
 (29b) \quad & \langle \nabla J(u), u - P_\Omega(u + \beta^{-1}\sigma v) \rangle \geq (s_N\beta^{-1}\sigma)^{-1} \|u + \beta^{-1}\sigma v_T - P_\Omega(u + \beta^{-1}\sigma v)\|^2 \\
 & - \langle P_T\nabla J(u), v_T \rangle \beta^{-1}\sigma - s_N^{-1}\kappa L_g\|x\| \|P_\Omega(u + \beta^{-1}\sigma v) - u\|\beta^{-1}\sigma.
 \end{aligned}$$

The estimates (26), (28), (29), and the parallelogram law imply that for all $u \in B(\bar{u}, \rho_1) \cap \Omega$, either $\sigma > \beta \min\{\alpha, \Delta\}$ or u is nonstationary and

$$\begin{aligned}
 (30a) \quad & c_1\|P_\Omega(u + \beta^{-1}\sigma v) - u\|^2\beta^{-1}\sigma \geq (1 - \delta)[\mu_3^{-1}\|u + \beta^{-1}\sigma v_T - P_\Omega(u + \beta^{-1}\sigma v)\|^2 \\
 & \quad + \mu_0\mu_1^{-2}\|v_T\|^2(\beta^{-1}\sigma)^2] \\
 & \geq c_2[\|u + \beta^{-1}\sigma v_T - P_\Omega(u + \beta^{-1}\sigma v)\|^2 + \|v_T\|^2(\beta^{-1}\sigma)^2] \\
 & \geq c_2\|P_\Omega(u + \beta^{-1}\sigma v) - u\|^2,
 \end{aligned}$$

where

$$(30b) \quad c_1 = \frac{1}{2}L_J + 2\mu_3\mu_0^{-1}\mu_2^{-2}(\mu_1 + \mu_3)\kappa L_g(1 + \beta^{-1}\alpha\kappa L_g)$$

and

$$(30c) \quad c_2 = \frac{1}{2}(1 - \delta)\min\{\mu_3^{-1}, \mu_0\mu_1^{-2}\}.$$

However, if u is nonstationary then Corollary 1 of Lemma 3 in [17] guarantees that $\|P_\Omega(u + \beta^{-1}\sigma v) - u\| > 0$ and thus

$$\sigma \geq \beta c_2 c_1^{-1} > 0.$$

Consequently, for all $u \in B(\bar{u}, \rho_1) \cap \Omega$,

$$\sigma \geq \min\{\alpha, \beta\Delta, \beta c_2 c_1^{-1}\} > 0.$$

A simpler but analogous argument applies when Q(i) is satisfied and σ is determined by (1), (2). As before, $\sigma = \alpha$ when u is stationary. If u is not stationary and $\sigma \leq \alpha\beta$ then (1m) implies a counterpart of (26) with $P_T\nabla J(u)$ replaced by $-P_{N^*}(-\nabla J(u))$. Moreover, with reference to (21), we obtain

$$\|P_\Omega(u + \beta^{-1}\sigma v) - \bar{u}\| \leq \|u - \bar{u}\| + \alpha(\mu_1 + \mu_3)\|P_{C(u)}(-\nabla J(u))\| + \alpha\mu_3\|\nabla J(u) - \nabla J(\bar{u})\|$$

with

$$\lim_{u \rightarrow \bar{u}} \|P_{C(u)}(-\nabla J(u))\| = 0$$

in place of (27). Hence there are numbers $\Delta > 0$, $\rho > 0$, and $\rho_1 \in (0, \rho]$ such that (42a), (42b) holds with (23). In view of (23), it can be seen that for all $u \in B(\bar{u}, \rho_1) \cap \Omega$, either $\sigma > \beta \min\{\alpha, \Delta\}$ or,

$$\begin{aligned} \langle \nabla J(u), u - P_\Omega(u + \beta^{-1}\sigma v) \rangle &\geq (s_N \beta^{-1}\sigma)^{-1} \|u + \beta^{-1}\sigma v_T - P_\Omega(u + \beta^{-1}\sigma v)\|^2 \\ &\quad + \langle P_{N^*}(-\nabla J(u)), v_T \rangle \beta^{-1}\sigma. \end{aligned}$$

Together, these estimates and the parallelogram law imply that for all $u \in B(\bar{u}, \rho_1) \cap \Omega$,

$$\sigma \geq \min\{\alpha, \beta\Delta, \beta c_2 c_1^{-1}\} > 0$$

with

$$c_1 = \frac{1}{2} L_J$$

and

$$c_2 = \frac{1}{2}(1 - \delta) \min\{\mu_3^{-1}, \mu_0 \mu_1^{-2}\}. \quad \square$$

THEOREM 3. *Assume that Ω is a closed convex set with a representation (4a) that satisfies the constraint qualification Q. Let \bar{u} be a nondegenerate stationary point and a uniformly proper local minimizer for J in Ω , and suppose that the gradients of J and the functions g_j in (4a) are Lipschitz continuous near \bar{u} in Ω . In addition, assume that \bar{u} is either a uniformly isolated stationary point, or more generally, a uniformly isolated zero of the nonstationarity measure $d(\cdot)$ in the active constraint manifold $\mathcal{A}(\bar{u})$. Then \bar{u} is an asymptotically stable fixed point for the SGP iteration (3), (4). Moreover, if Q(i) holds then \bar{u} is an asymptotically stable fixed point for (1), (2) as well. In either case, every associated SGP sequence $\{u^i\}$ that converges to \bar{u} is eventually confined to the active constraint manifold $\mathcal{A}(\bar{u})$.*

Proof. Suppose that Q(ii) holds. Then by Lemma 5, \bar{u} is a stable fixed point for (3), (4). Suppose that \bar{u} is a uniformly isolated zero of $d(\cdot)$ in $\mathcal{A}(\bar{u})$. Then for some $\rho_1 > 0$, some nondecreasing positive-definite real function $a(\cdot)$, and all u ,

$$(31) \quad u \in B(\bar{u}, \rho_1) \cap \mathcal{A}(\bar{u}) \Rightarrow d(u) \geq a(\|u - \bar{u}\|).$$

According to Lemma 7, there are numbers $\rho_2 \in (0, \rho_1]$, $\alpha_1 > 0$ and $\alpha_2 > 0$ such that for all u ,

$$u \in B(\bar{u}, \rho_2) \cap \Omega \text{ and } (3) - (4) \Rightarrow \alpha_2 \geq \sigma s_N \geq \alpha_1.$$

With reference to the proof of Lemma 5, note that for (3), (4),

$$(32a) \quad P_\Omega(u + \sigma v) = P_\Omega(\bar{u} - \sigma s_N \nabla J(\bar{u}) + h(u))$$

with

$$(32b) \quad \|h(u)\| \leq \|u - \bar{u}\| + \alpha(\mu_1 + \mu_3) \|P_{T(u)} \nabla J(u)\| + \alpha \mu_3 \|\nabla J(u) - \nabla J(\bar{u})\|;$$

therefore,

$$(32c) \quad \lim_{\substack{u \rightarrow \bar{u} \\ u \in \Omega}} h(u) = 0.$$

Hence by Lemma 3, there is a $\rho_3 \in (0, \rho_2]$ such that for all u ,

$$u \in B(\bar{u}, \rho_3) \cap \Omega \text{ and } (3) - (4) \Rightarrow P_\Omega(u + \sigma v) \in \mathcal{A}(\bar{u}).$$

Furthermore, since \bar{u} is stable, there is a $\rho_4 \in (0, \rho_3]$ such that if $u^1 \in B(\bar{u}, \rho_4) \cap \Omega$ and $\{u^i\}$ is generated by (3), (4), then $u^i \in B(\bar{u}, \rho_3) \cap \Omega$ for all $i \geq 1$; therefore $u^i \in B(\bar{u}, \rho_3) \cap \mathcal{A}(\bar{u})$ for all $i \geq 2$. For all such sequences $\{u^i\}$, the corresponding step lengths σ^i are bounded away from zero and Lemma 4 in [17] implies that

$$\lim_{i \rightarrow \infty} d(u^i) = 0;$$

therefore,

$$\lim_{i \rightarrow \infty} \|u^i - \bar{u}\| = 0$$

in view of (31).

Now suppose that Q(i) holds and \bar{u} is a uniformly isolated zero of $d(\cdot)$ in $\mathcal{A}(\bar{u}) = \bar{u} + K(\bar{u})^\perp$. For (1), (2), the estimate (32) is replaced by

$$P_\Omega(u + \sigma v) = P_\Omega(\bar{u} - \sigma s_N \nabla J(\bar{u}) + h(u))$$

with

$$\|h(u)\| \leq \|u - \bar{u}\| + \alpha(\mu_1 + \mu_3) \|P_{C(u)}(-\nabla J(u))\| + \alpha\mu_3 \|\nabla J(u) - \nabla J(\bar{u})\|;$$

therefore,

$$\lim_{u \rightarrow \bar{u}} h(u) = 0.$$

Lemmas 3, 5, and 7 are then used as before to show that \bar{u} is asymptotically stable for (1), (2) and hence (3), (4) (see Note 1 in [17]), and that every corresponding SGP sequence $\{u^i\}$ which converges to \bar{u} , eventually enters and remains in $\mathcal{A}(\bar{u})$. \square

THEOREM 4. *Assume that Ω is a closed convex set with a representation (4a) that satisfies the constraint qualification Q. Let \bar{u} be a nonsingular local minimizer for J in Ω . Then \bar{u} is an asymptotically stable fixed point for the SGP iteration (3), (4). Moreover, if Q(i) holds then \bar{u} is an asymptotically stable fixed point for (1), (2) as well. In either case, every associated SGP sequence $\{u^i\}$ that converges to \bar{u} is eventually confined to the active constraint manifold $\mathcal{A}(\bar{u})$. Furthermore, if $\{u^i\}$ converges to \bar{u} , and if $u^i \neq \bar{u}$ for all i , then for some $r \in [0, 1)$ and some positive real sequence $\{\rho_i\}$,*

$$(33a) \quad \limsup_{i \rightarrow \infty} \frac{J(u^{i+1}) - J(\bar{u})}{J(u^i) - J(\bar{u})} = r$$

and

$$(33b) \quad \|u^i - \bar{u}\|^2 \leq \rho_i, \quad i = 1, 2, \dots$$

with

$$(33c) \quad \limsup_{i \rightarrow \infty} \frac{\rho_{i+1}}{\rho_i} = r.$$

Proof. According to Definition 3 and the estimate (11a) in Lemma 2, \bar{u} is a nondegenerate stationary point and a uniformly proper local minimizer; moreover, the gradients of J and the functions g_j are Lipschitz continuous near \bar{u} in Ω . All of our claims except (33) will then follow from Theorem 3 if it can be shown that \bar{u} is also a uniformly isolated zero of $d(\cdot)$ in $\mathcal{A}(\bar{u})$. When Q(i) holds, this can be proved with Lemma 3.1 in [8], or more directly as follows. Note that by Lemma 4,

$$P_\Omega(u - \nabla J(u)) \in \mathcal{A}(\bar{u}) = \bar{u} + T_0(\bar{u});$$

therefore,

$$[u - \nabla J(u) - P_\Omega(u - \nabla J(u))] \in K(P_\Omega(u - \nabla J(u)) = K(\bar{u}) \subset T_0(\bar{u})^\perp$$

for u sufficiently near \bar{u} in Ω ; moreover, if u is also in $\mathcal{A}(\bar{u})$, then

$$u - P_\Omega(u - \nabla J(u)) \in T_0(\bar{u}).$$

Consequently, for u near \bar{u} in $\mathcal{A}(\bar{u})$ we have

$$(u - P_\Omega(u - \nabla J(u)) - P_{T_0(\bar{u})} \nabla J(u)) \in T_0(\bar{u}) \cap T_0(\bar{u})^\perp$$

and hence

$$u - P_\Omega(u - \nabla J(u)) - P_{T_0(\bar{u})} \nabla J(u) = 0;$$

therefore,

$$d(u) = \|P_{T_0(\bar{u})} \nabla J(u)\| = \|P_{T_0(u)} \nabla J(u)\|.$$

The estimate (11b) in Lemma 2 now implies that \bar{u} is a uniformly isolated zero of $d(\cdot)$ in $\mathcal{A}(\bar{u})$. Similarly, if Q(ii) holds, then Lemma 2 in [9] asserts that

$$d(u) \geq c \|P_{T_0(u)} \nabla J(u)\|$$

for some $c > 0$ and all u near \bar{u} in $\mathcal{A}(\bar{u})$, in which case (11b) insures once again that \bar{u} is a uniformly isolated zero of $d(\cdot)$ in $\mathcal{A}(\bar{u})$.

To prove (33) suppose first that Q(ii) holds. Then, according to Lemma 3 in [9],

$$(34) \quad \langle P_{T_0(u)} \nabla J(u), u - \bar{u} \rangle \geq J(u) - J(\bar{u})$$

for u sufficiently near \bar{u} in $\mathcal{A}(\bar{u})$. In view of (11a) this implies that

$$\|P_{T_0(u)} \nabla J(u)\|^2 \geq c_1 (J(u) - J(\bar{u}))$$

for u near \bar{u} in $\mathcal{A}(\bar{u})$. Hence if $\{u^i\}$ is generated by (3), (4) and converges to \bar{u} , the inequalities

$$J(u^i) - J(u^{i+1}) \geq \delta \mu_0 c_1 (J(u^i) - J(\bar{u}))$$

and

$$J(u^i) - J(\bar{u}) \geq c_1 \|u^i - \bar{u}\|^2$$

must hold eventually. This establishes (33) for (3), (4). On the other hand, suppose that Q(i) holds. Since \bar{u} is nonsingular, the restriction of J to $\mathcal{A}(\bar{u}) = \bar{u} + T_0(\bar{u})$ is convex near \bar{u} ; therefore, (34) is satisfied once again near \bar{u} in $\mathcal{A}(\bar{u})$. A repetition of the argument following (34) will prove (33) for (1), (2) and hence for (3), (4) (see Note 1 in [17]). \square

Note 7. When Q(i) holds, Theorem 2 and Lemma 3.2 in [8] can be used to establish sublinear convergence rate estimates for (1), (2) and (3), (4) near certain asymptotically stable singular minimizers \bar{u} .

4. Newtonian convergence acceleration. The linear convergence rate results in Theorem 4 are “worst case” estimates for general SGP iterations (1), (2) and (3), (4) near nonsingular minimizers; in particular, these estimates apply to the original unscaled Goldstein–Levitin–Polyak GP iteration obtained when $s_N = 1$ and $N = \mathcal{U}$ in (1). The analysis in this section now focuses more narrowly on SGP iterations (3), (4) that employ the smallest admissible subspaces $N = N(u)$, and hence the largest admissible scaling subspaces $T = T(u) = N(u)^\perp$. On the face of it, these algorithms offer the greatest latitude of choice for local convergence acceleration within the scheme (3), (4), and it is shown below that the special SGP iteration (3)–(6) is indeed locally superlinearly convergent to nonsingular minimizers in closed convex sets with representations (4a) satisfying Q. Moreover, when Q(i) holds, a variant of Proposition 3 in [15] is obtained by demonstrating that, near nondegenerate stationary points \bar{u} at which ∇J is continuous, every SGP iteration (1), (2) with $N = C(u)^*$ is locally equivalent to some SGP iteration (3), (4) with $N = N(u)$. This last result will be proved first.

LEMMA 8. *Assume that Ω is a polyhedral convex set with a representation (4a) that satisfies the constraint qualification Q(i). Let \bar{u} be a nondegenerate stationary point for J in Ω and suppose that J is continuously Fréchet differentiable at \bar{u} . Then there is a $\rho > 0$ such that in $B(\bar{u}, \rho) \cap \Omega$, every SGP map $u \rightarrow P_\Omega(u + \sigma v)$ satisfying (1), (2) with $N = C(u)^*$ coincides with some SGP map $u \rightarrow P_\Omega(u + \sigma v)$ satisfying (3), (4) with $N = N(u)$.*

Proof. For $u \in \Omega$, let

$$x(u) \triangleq P_{C(u)}(-\nabla J(u)), \quad y(u) \triangleq P_{C(u)^*}(-\nabla J(u)).$$

It suffices to show that for some $\rho > 0$ and all $u \in B(\bar{u}, \rho) \cap \Omega$,

$$(35a) \quad y(u)^\perp \cap C(u) = T(u),$$

$$(35b) \quad x(u) = -P_{T(u)}\nabla J(u),$$

$$(35c) \quad y(u) = -P_{N(u)}\nabla J(u).$$

According to Lemma 4, there is a $\rho_1 > 0$ such that for all $u \in B(\bar{u}, \rho_1) \cap \Omega$,

$$(36) \quad C(u)^* = K(\bar{u}) \quad \text{and} \quad N(u) = [K(\bar{u})].$$

Hence (35a) will follow if it can be shown that for some $\rho \in (0, \rho_1]$ and all $u \in B(\bar{u}, \rho) \cap \Omega$

$$y(u)^\perp \cap K(\bar{u})^* = K(\bar{u})^\perp.$$

Note that $K(\bar{u})^\perp \subset K(\bar{u})^*$. Furthermore, $y(u) \in K(\bar{u})$ for all $u \in B(\bar{u}, \rho_1) \cap \Omega$, in view of (36). Hence, for all $u \in B(\bar{u}, \rho_1) \cap \Omega$,

$$K(u)^\perp \subset y(u)^\perp \cap K(\bar{u})^*.$$

To prove that the reverse inclusion holds near \bar{u} in Ω , let

$$h(u) = P_{K(\bar{u})}(-\nabla J(u)) - P_{K(\bar{u})}(-\nabla J(\bar{u})) \in [K(\bar{u})]$$

and observe that

$$\|h(u)\| \leq \|\nabla J(u) - \nabla J(\bar{u})\|.$$

Since $-\nabla J(\bar{u}) \in \text{ri } K(\bar{u})$, and $\nabla J(\cdot)$ is continuous at \bar{u} , there is a $\rho \in (0, \rho_1]$ such that for all $u \in B(\bar{u}, \rho) \cap \Omega$

$$P_{K(\bar{u})}(-\nabla J(u)) \in \text{ri } K(\bar{u}),$$

i.e., for some $\epsilon > 0$ and all η ,

$$(37) \quad \eta \in [K(\bar{u})] \quad \text{and} \quad \|\eta\| \leq \epsilon \Rightarrow (P_{K(\bar{u})}(-\nabla J(u)) + \eta) \in K(\bar{u}).$$

Now suppose that $w \in y(u)^\perp \cap K(\bar{u})^*$. In view of (37) it can be seen that for all $u \in B(\bar{u}, \rho) \cap \Omega$ and all $\eta \in [K(\bar{u})] \cap B(0, \epsilon)$,

$$\langle \eta, w \rangle = \langle P_{K(\bar{u})}(-\nabla J(u)) + \eta, w \rangle \leq 0.$$

Since $[K(\bar{u})]$ is a subspace, it follows that $w \in K(\bar{u})^\perp$. Consequently, for all $u \in B(\bar{u}, \rho) \cap \Omega$,

$$y(u)^\perp \cap K(\bar{u})^* \subset K(u)^\perp.$$

This establishes (35a). Next, observe that

$$(38a) \quad -\nabla J(u) = x(u) + y(u)$$

and

$$(38b) \quad \langle x(u), y(u) \rangle = 0$$

(cf. [20, Lemma 2.2]). According to (35a) and (38b), we therefore have

$$x(u) \in T(u) = y(u)^\perp \cap C(u) \subset C(u)$$

for all $u \in B(\bar{u}, \rho) \cap \Omega$. This proves (35b). Finally, conditions (35b) and (38a) yield

$$\begin{aligned} y(u) &= -\nabla J(u) - x(u) \\ &= -\nabla J(u) + P_{T(u)} \nabla J(u) \\ &= -P_{N(u)} \nabla J(u) \end{aligned}$$

for all $u \in B(\bar{u}, \rho) \cap \Omega$.

THEOREM 5. *Let \bar{u} be a nonsingular local minimizer for J in a polyhedral convex set Ω with a representation (4a) satisfying the constraint qualification Q(i). Then there are positive numbers ρ and $\bar{\mu}_1 \geq \bar{\mu}_0$ such that for all $u \in B(\bar{u}, \rho) \cap \Omega$, the operators $P_{T(u)} \nabla^2 J(u)|_{T(u)}$ are bijective and the corresponding inverse operators,*

$$S_T(u) = (P_{T(u)} \nabla^2 J(u)|_{T(u)})^{-1},$$

in (5) satisfy the conditions

$$(39a) \quad \|S_T(u)\| \leq \bar{\mu}_1$$

and

$$(39b) \quad \inf_{w \in T(u)} \langle w, S_T(u)w \rangle \geq \bar{\mu}_0 \|w\|^2.$$

Furthermore, suppose that $\{u^i\}$ is generated by an SGP iteration (3), (4), with $\mu_1 \geq \bar{\mu}_1 \geq \bar{\mu}_0 \geq \mu_0$, $\delta \in (0, \frac{1}{2})$, and $\alpha = 1$, and assume that for all i ,

$$(40) \quad u^i \in B(\bar{u}, \rho) \cap \Omega \Rightarrow N^i = N(u^i) \quad \text{and} \quad S_T^i = S_T(u^i).$$

Then $\{u^i\}$ converges to \bar{u} , provided u^1 is sufficiently near \bar{u} in Ω ; moreover, if $\{u^i\}$ converges to \bar{u} , then eventually,

$$(41a) \quad u^i \in \mathcal{A}(\bar{u}) = \bar{u} + T_0(\bar{u}),$$

$$(41b) \quad u^{i+1} = u^i + \sigma^i v_T^i,$$

and

$$\sigma^i = \max s$$

subject to

$$s \in \{1, \beta, \beta^2, \dots\}$$

and

$$(41c) \quad J(u^i) - J(P_\Omega(u^i + s v_T^i)) \geq -\delta \langle P_{T_0(\bar{u})} \nabla J(u^i), v_T^i \rangle s,$$

where

$$(41d) \quad v_T^i = -(P_{T_0(\bar{u})} \nabla^2 J(u^i)|_{T_0(\bar{u})})^{-1} P_{T_0(\bar{u})} \nabla J(u^i).$$

Accordingly,

$$(42a) \quad \sigma^i = 1$$

for sufficiently large i , and $\{u^i\}$ converges q -superlinearly to \bar{u} , i.e., either $u^i = \bar{u}$ eventually, or

$$(42b) \quad \lim_{i \rightarrow \infty} \frac{\|u^{i+1} - \bar{u}\|}{\|u^i - \bar{u}\|} = 0.$$

Proof. By Definition 3, the reduced Hessian $P_{T_0(\bar{u})} \nabla^2 J(\bar{u})|_{T_0(\bar{u})}$ is self adjoint and coercive, and $\nabla^2 J(\cdot)$ is continuous at \bar{u} . By Lemma 4, $N(u) = N_0(\bar{u})$ and $T(u) = T_0(\bar{u})$ near \bar{u} in Ω . Hence the operators $P_{T(u)} \nabla^2 J(u)|_{T(u)}$ and their inverses are bounded and coercive uniformly near \bar{u} in Ω . By Theorem 4, all sequences $\{u^i\}$ generated by (3), (4) must converge to \bar{u} from nearby starting points u^1 in Ω , and must eventually enter and remain within $\mathcal{A}(\bar{u}) = \bar{u} + T_0(\bar{u})$, as stated in (41a). For large i , we therefore have

$$u^i - u^{i+1} \in T_0(\bar{u}), \quad [u^i + \sigma^i v^i - u^{i+1}] \in K(u^{i+1}) = K(\bar{u}) \subset N_0(\bar{u}).$$

Thus, (40) insures that for large i

$$(u^i - u^{i+1} + \sigma^i v_T^i) \in T_0(\bar{u}) \cap N_0(\bar{u}) = \{0\}.$$

This proves (41b) and hence (41c). The remaining assertions are obtained from Proposition 1.15 in [21], after noting that (41b), (41c) amounts to an unconstrained relaxed Newton iteration with Armijo steps. \square

Note 8. Assertions (41a)–(41c) and (42) continue to hold if the Newtonian scaling restriction in (40) is replaced by the weaker quasi-Newton condition

$$(43) \quad \lim_{i \rightarrow \infty} \frac{\|(S_T^i - S_T(u^i))P_{T(u^i)}\nabla J(u^i)\|}{\|P_{T(u^i)}\nabla J(u^i)\|} = 0.$$

(See Proposition 1.15 in [21].)

Note 9. An application of Lemma 8 yields a counterpart of Theorem 5 for SGP iterations (1), (2) satisfying $N^i = C(u^i)^*$ in place of $N^i = N(u^i)$; this result may be viewed as a supplement to the Newtonian convergence acceleration consequences of Proposition 3 in [15] (note that in [15], \bar{u} is *assumed* to be a subsequential limit of the sequence $\{u^i\}$).

In nonpolyhedral convex Ω , the normal scaling parameters s_N have a more important bearing on convergence acceleration than heretofore. Previously, the numbers s_N^i merely had to remain bounded away from 0 and ∞ , but now they must satisfy additional restrictions to insure that near nonsingular minimizers, the Newtonian scaling operators $S_T(u)$ in (5) satisfy conditions (3i), (3j), that the step length rule (3) admits $\sigma = 1$ when $\delta \in (0, \frac{1}{2})$ and $\alpha = 1$, and hence that (3)–(5) reduce to the superlinearly convergent NP iteration in [16]. All these requirements are met by the rule (6).

THEOREM 6. *Let \bar{u} be a nonsingular local minimizer for J in a closed convex set Ω with a representation (4a) satisfying the constraint qualification Q(ii). Then there are positive numbers ρ , $\bar{\mu}_3 \geq \bar{\mu}_2$, $\bar{\mu}_1 \geq \bar{\mu}_0$ such that for all $u \in B(\bar{u}, \rho) \cap \Omega$ the operators $P_{T(u)}\nabla^2 J(u)|_{T(u)}$ are bijective, and the corresponding operators $S_T(u)$ and scaling parameters $s_N(u)$ in (5), (6) are well defined and satisfy the conditions*

$$(44a) \quad \bar{\mu}_3 \geq s_N(u) \geq \bar{\mu}_2,$$

$$(44b) \quad \langle P_{T(u)}\nabla J(u), S_T(u)P_{T(u)}\nabla J(u) \rangle \geq \bar{\mu}_0 \|P_{T(u)}\nabla J(u)\|^2,$$

$$(44c) \quad \|S_T(u)\| \leq \bar{\mu}_1.$$

Furthermore, suppose that $\{u^i\}$ is generated by an SGP iteration (3), (4) with $\mu_1 \geq \bar{\mu}_1 \geq \bar{\mu}_0 \geq \mu_0$, $\mu_3 \geq \bar{\mu}_3 \geq \bar{\mu}_2 \geq \mu_2$, $\delta \in (0, \frac{1}{2})$, and $\alpha = 1$, and assume that for all i

$$(45) \quad u^i \in B(\bar{u}, \rho) \cap \Omega \Rightarrow N^i = N(u^i), \quad s_N^i = s_N(u^i) \quad \text{and} \quad S_T^i = S_T(u^i)$$

Then $\{u^i\}$ converges to \bar{u} provided u^1 is sufficiently close to \bar{u} ; moreover, if $\{u^i\}$ converges to \bar{u} , then eventually

$$(46a) \quad u^i \in \mathcal{A}(\bar{u}),$$

$$(46b) \quad \sigma^i = 1,$$

and hence $\{u^i\}$ converges q -superlinearly to \bar{u} .

Proof. It is shown in §§3 and 4 of [16] that the multiplier function $\lambda(\cdot)$ in (5) is continuous at \bar{u} , that the associated operators $L_T(u)$ and their inverses are uniformly bounded and coercive near \bar{u} , and that

$$(47) \quad \lim_{\substack{u \rightarrow \bar{u} \\ u \in \Omega}} P_{T(u)}\nabla J(u) = 0.$$

Since the operators $\Lambda_T(u)$ are also uniformly bounded near \bar{u} , it can be seen that conditions (5) and (6) uniquely define scale factors $s_N(u)$ and scaling operators $S_T(u)$

satisfying (44). By Theorem 4, all sequences $\{u^i\}$ generated by (3), (4) must converge to \bar{u} from nearby starting points u^1 , and eventually satisfy (46a). Condition (46b) will now follow from (3k) if it can be shown that for $u \in \mathcal{A}(\bar{u})$ sufficiently near \bar{u} ,

$$(48a) \quad J(u) - J(\phi(u, 1)) - \delta\{s_N(u)^{-1}\|u - S_T(u)P_{T(u)}\nabla J(u) - \phi(u, 1)\|^2 + \langle P_{T(u)}\nabla J(u), S_T(u)P_{T(u)}\nabla J(u) \rangle\} \geq 0,$$

where

$$(48b) \quad \phi(u, s) = P_\Omega(u - s s_N(u)P_{N(u)}\nabla J(u) - s S_T(u)P_{T(u)}\nabla J(u)).$$

With reference to (32), note that

$$\phi(u, 1) = P_\Omega(\bar{u} - s_N(u)\nabla J(\bar{u}) + h(u))$$

with $s_N(u)$ bounded away from 0 and ∞ near \bar{u} , and

$$\lim_{\substack{u \rightarrow \bar{u} \\ u \in \Omega}} h(u) = 0.$$

Therefore, for $u \in \Omega$ near \bar{u} ,

$$\phi(u, 1) \in \mathcal{A}(\bar{u}),$$

by Lemma 3. Consequently, for $u \in \mathcal{A}(\bar{u})$ near \bar{u} , and for all i ,

$$\mathcal{J}(u) = \mathcal{J}_0(u) = \mathcal{J}_0(\phi(u, 1)) = \mathcal{J}_0(\bar{u})$$

and

$$\begin{aligned} i \in \mathcal{J}(u) \Rightarrow 0 = g_i(u) - g_i(\phi(u, 1)) &= \langle \nabla g_i(u), u - \phi(u, 1) \rangle \\ &\quad - \frac{1}{2} \langle u - \phi(u, 1), \nabla^2 g_i(u)(u - \phi(u, 1)) \rangle \\ &\quad + o(\|u - \phi(u, 1)\|^2). \end{aligned}$$

In view of (5), this yields

$$0 = -\langle P_{N(u)}\nabla J(u), u - \phi(u, 1) \rangle - \frac{1}{2} \langle u - \phi(u, 1), \Lambda(u)(u - \phi(u, 1)) \rangle + o(\|u - \phi(u, 1)\|^2);$$

therefore,

$$(49) \quad \begin{aligned} J(u) - J(\phi(u, 1)) &= \langle P_{T(u)}\nabla J(u), u - \phi(u, 1) \rangle \\ &\quad - \frac{1}{2} \langle u - \phi(u, 1), L(u)(u - \phi(u, 1)) \rangle + o(\|u - \phi(u, 1)\|^2) \end{aligned}$$

as $u \rightarrow \bar{u}$ with $u \in \mathcal{A}(\bar{u})$. In addition, the analysis in §4 of [16] shows that for $u \in \mathcal{A}(\bar{u})$ near \bar{u} ,

$$(50a) \quad \phi(u, 1) - \bar{u} = o(\|u - \bar{u}\|),$$

$$(50b) \quad u - \bar{u} = P_{T(u)}(u - \bar{u}) + o(\|u - \bar{u}\|),$$

and

$$(50c) \quad P_{T(u)}\nabla J(u) = L_T(u)^{-1}P_{T(u)}(u - \bar{u}) + o(\|u - \bar{u}\|);$$

therefore,

$$(51) \quad u - \phi(u, 1) = L_T(u)^{-1} P_{T(u)} \nabla J(u) + o(\|P_{T(u)} \nabla J(u)\|)$$

as $u \rightarrow \bar{u}$ with $u \in \mathcal{A}(\bar{u})$. (In particular, see (4.13), (4.23), Lemma 4.2, Corollary 4.2.1, Lemma 4.3, and the remark concerning $s_N(u)$ at the beginning of §4 in [16].) In view of (49) and (51), it can now be seen that the left side of (48a) is bounded below by

$$\begin{aligned} & \left(\frac{1}{2} - \delta \right) \langle P_{T(u)} \nabla J(u), L_T(u)^{-1} P_{T(u)} \nabla J(u) \rangle - \delta s_N(u) \{ \|\Lambda_T(u) L_T(u)^{-1} P_{T(u)} \nabla J(u)\|^2 \\ & \quad + |\langle P_{T(u)}, \Lambda_T(u) L_T(u)^{-1} P_{T(u)} \nabla J(u) \rangle| \} + o(\|P_{T(u)} \nabla J(u)\|^2) \\ & \geq \frac{(1 - 2\delta)^2}{2} \langle P_{T(u)} \nabla J(u), L_T(u)^{-1} P_{T(u)} \nabla J(u) \rangle + o(\|P_{T(u)} \nabla J(u)\|^2) \geq 0 \end{aligned}$$

as $u \rightarrow \bar{u}$ with $u \in \mathcal{A}(\bar{u})$. This estimate and (46a) establish (46b), and q -superlinear convergence of $\{u^i\}$ then follows from (50a). \square

Note 10. If the Hessians of J and the functions g_j are Lipschitz continuous near a nonsingular minimizer \bar{u} , then the q -superlinear convergence rate estimate in Theorems 5 and 6 can be replaced with a sharper q -quadratic convergence rate estimate (see Note 4.2 in [16]).

Note 11. An examination of (4.13), (4.23) and the related analysis in §4 of [16] will show that the conclusions in Theorem 6 also remain valid if the Newtonian scaling restriction in (45) is replaced by the quasi-Newton condition (43).

REFERENCES

- [1.] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.
- [2.] E. S. LEVITIN AND B. T. POLJAK, *Constrained optimization methods*, USSR Comp. Math. Phys., 6 (1966), pp. 1–50.
- [3.] V. F. DEMYANOV AND A. M. RUBINOV, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1970.
- [4.] G. P. MCCORMICK AND R. A. TAPIA, *The gradient projection method under mild differentiability conditions*, SIAM J. Control, 10 (1972), pp. 93–98.
- [5.] D. P. BERTSEKAS, *On the Goldstein–Levitin–Poljak gradient projection method*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 174–184.
- [6.] B. N. PSHENICHNY AND YU. M. DANILOV, *Numerical Methods in Extremal Problems*, MIR, Moscow, Russia, 1978.
- [7.] J. C. DUNN, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1980), pp. 368–400.
- [8.] ———, *On the convergence of projected gradient processes to singular attractors*, J. Optim. Theory Appl., 55 (1987), pp. 203–215.
- [9.] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Appl. Math. Optim., 17 (1988), pp. 103–119.
- [10.] J. V. BURKE AND J. J. MORÉ, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.
- [11.] P. H. CALAMAI AND J. J. MORÉ, *Projected Gradient Methods for Linearly Constrained Problems*, Tech. Memo. MCM-73, Argonne National Laboratory, Argonne, IL, May, 1986.
- [12.] W. A. GRUVER AND E. W. SACHS, *Algorithmic Methods in Optimal Control*, Pitman, Boston, MA, 1980.
- [13.] J. C. DUNN, *Gradient projection methods for systems optimization problems*, in Control and Dynamic Systems, Vol. 29, C. T. Leondes, ed., Academic Press, Orlando, FL, 1988.

- [14] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control, 20 (1982), pp. 221–246.
- [15] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained minimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.
- [16] J. C. DUNN, *A projected Newton method for minimization problems with nonlinear inequality constraints*, Numer. Math., 53 (1988), pp. 377–409.
- [17] ———, *A subspace decomposition principle for scaled gradient projection methods: Global theory*, SIAM J. Control Optim., 29 (1991), pp. 1150–1175.
- [18] ———, *Asymptotic decay rates from the growth properties of Liapunov functions near singular attractors*, J. Math. Anal. Appl., 125 (1987), pp. 6–21.
- [19] V. M. ALEKSEEV, V. M. TIKHOMIROV, AND S. V. FOMIN, *Optimal Control*, Plenum Press, New York, 1987.
- [20] E. H. ZARANTONELLO, *Projections on convex sets in Hilbert space, and spectral theory*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971.
- [21] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

TWO GENERALIZATIONS OF A THEOREM OF ARROW, BARANKIN, AND BLACKWELL*

RICHARD J. GALLAGHER[†] AND OSSAMA A. SALEH[‡]

Abstract. In 1953, Arrow, Barankin, and Blackwell proved that if R^n is equipped with its natural ordering and if A is a closed convex subset of R^n , then the set of points in A that can be supported by strictly positive linear functionals is dense in the set of all efficient (maximal) points of A . In this note two generalizations of this result are given. The first of these is in the setting of a dual system and requires relatively weak assumptions on the ordering cone but a rather strong compactness assumption on the set A . The second generalization, which is in the setting of a locally convex space, relaxes the compactness assumption on the set A but demands more stringent assumptions on the ordering cone. This second result was recently obtained by Petschke for normed spaces [M. Petschke, “On a theorem of Arrow, Barankin and Blackwell”, *SIAM J. Control Optim.*, 28 (1990), pp. 395–401]. The proof given here is substantially different from that given by Petschke.

Key words. vector optimization, scalarization, proper efficiency

AMS(MOS) subject classifications. 90C31, 52A07

1. Introduction. In 1953 Arrow, Barankin, and Blackwell [1] proved that if R^n is equipped with its natural ordering and if A is a closed convex subset of R^n , then the set of points in A that can be supported by strictly positive linear functionals is dense in the set of all efficient (maximal) points of A . (Here, a point $a_0 \in A$ is said to be supported by a linear functional f if $f(a_0) \geq f(a)$ for all $a \in A$.) This theorem has important implications in mathematical economics and multiple objective optimization. In particular, in mathematical economics a strictly positive linear functional p can be regarded as a pricing system, and a point $a_0 \in A$ supported by p represents an efficient allocation of resources that can be sustained by the pricing system p . Hence, the theorem states that under appropriate conditions, “nearly every” efficient allocation of resources can be sustained by a suitable pricing system. In the case of multiple objective optimization where one is interested in identifying the set of all efficient points of a given set A , the theorem implies that “nearly every” efficient point of A can be found as the solution of a suitable single-objective real-valued optimization problem.

Over the past 25 years, several authors have generalized the Arrow–Barankin–Blackwell theorem in various ways. Hartley [12] and Bitran and Magnanti [2] showed that the theorem remains valid if R^n is partially ordered by an arbitrary closed pointed convex cone. Three mathematical economists, Radner [24], Majumdar [19], and Peleg [21], obtained density results in the infinite dimensional normed vector lattices ℓ^∞ , L^∞ and ℓ^p , $1 \leq p \leq \infty$, respectively. Chichilnisky and Kalman [6] proved a density theorem in a Hilbert space setting; and Salz [26], Borwein [3], [4], Jahn [13], Dauer

*Received by the editors January 21, 1991; accepted for publication (in revised form) November 12, 1991.

[†]Department of Mathematical Sciences, University of Montana, Missoula, Montana 59812. The research of this author was supported in part by funds provided by the Provident Chair of Excellence in Applied Mathematics at the University of Tennessee at Chattanooga, and by a grant from the Montanans on a New Trac for Science program.

[‡]Department of Mathematics, University of Tennessee at Chattanooga, Chattanooga, Tennessee 37403.

and Gallagher [9], and Petschke [23] proved density results in general normed spaces. Three other papers that should be mentioned are those of Peleg [20], which includes a constructive proof of the original theorem of Arrow, Barankin, and Blackwell; and DaCunha and Polak [8] and Durier [10], which include related density theorems for constrained multiple objective problems having a finite number of objectives.

A brief examination of the various hypotheses used in [3], [4], [9], [13], [23], and [26] to obtain density results in general normed spaces, reveals that there is a trade-off between restrictions placed on the ordering cone and restrictions placed on the set A . That is, existing density results either require strong assumptions on the ordering cone and relatively weak assumptions on the set A , or relatively weak assumptions on the ordering cone and strong assumptions on the set A . However, no counterexamples appear in the literature that demonstrate that such trade-offs are necessary. Hence, further study is warranted.

It is the purpose of this paper to present two density theorems, in the most general settings known to the authors, which exemplify the above-mentioned trade-offs. It is hoped that the unified treatment given here will provide some insight and direction towards either obtaining density results with less restrictive hypotheses, or showing that the known results are, in some sense, the best that can be obtained.

The first theorem (Theorem 2.6) generalizes the density results in [3], [4], [9], and [24] to the setting of a dual system. Although the technique of proof relies on a standard argument, abstracting the problem to this setting emphasizes the precise properties that the ordering cone must satisfy in order for the argument to remain valid. More important, by Corollary 2.8 it follows that the density result of Borwein [3, Thm. 2] remains true without requiring the cone to have a weakly compact base. Indeed, the result only requires the cone to be closed and convex and to admit strictly positive continuous linear functionals. Borwein himself noted this in a later paper [4, Thm. 5], but made no comment on the proof. Since two recent papers [13], [23] quote only the former result, it seems that even in a normed space setting, Corollary 2.8 is not widely known.

The second density theorem presented in this paper (Theorem 3.1) extends a recent result of Petschke [23] to the setting of a locally convex space. Petschke has shown that in a normed space the Arrow–Barankin–Blackwell theorem holds provided the ordering cone has a closed bounded base and the set A is weakly compact and convex. The proof given by Petschke is somewhat lengthy and depends strongly on the norm structure. In Theorem 3.1 it is shown that the result is even true in a locally convex space. The proof given is quite short and relies only on standard separation arguments.

The paper is organized as follows. In §2 we introduce the notion of a D-cone, demonstrate that the properties of such a cone are relatively easy to satisfy in locally bounded spaces (in particular, in normed spaces), and establish Theorem 2.6 which assumes that the ordering is induced by a D-cone. In §3 we review the notion of a base for a cone, discuss the relationship between D-cones and cones with a closed bounded base, and establish Theorem 3.1 which assumes that the ordering is induced by a cone with a closed bounded base. In both §§2 and 3 several corollaries of the main results are given. Some concluding remarks and an open question are given in §4.

2. Density for a space ordered by a D-cone. We begin by reviewing the definition of a dual system. For a comprehensive treatment of this concept, the reader is referred to Schaefer [27].

Let \mathcal{X} and \mathcal{F} be vector spaces over R , and let $\langle \cdot, \cdot \rangle$ be a bilinear form on $\mathcal{X} \times \mathcal{F}$ satisfying the separation axioms (2.1) and (2.2) below.

$$(2.1) \quad \text{If } \langle x_0, f \rangle = 0 \text{ for all } f \in \mathcal{F}, \text{ then } x_0 = \theta.$$

$$(2.2) \quad \text{If } \langle x, f_0 \rangle = 0 \text{ for all } x \in \mathcal{X}, \text{ then } f_0 = \theta.$$

(The symbol θ denotes the zero vector in the appropriate space.) Then the triple $(\mathcal{X}, \mathcal{F}, \langle \cdot, \cdot \rangle)$ is called a *dual system* over R , and is usually denoted more briefly as $\langle \mathcal{X}, \mathcal{F} \rangle$. The prototype example of a dual system is the case when \mathcal{X} is a normed space and \mathcal{F} is the topological dual \mathcal{X}^* of \mathcal{X} . In this case the bilinear form $\langle x, f \rangle = f(x)$ defines a dual system.

A subset $K \subseteq \mathcal{X}$ is said to be a *pointed convex cone* if K is convex, $\lambda K \subseteq K$ for all $\lambda \geq 0$, and $K \cap (-K) = \{\theta\}$.

Throughout this section, let $\langle \mathcal{X}, \mathcal{F} \rangle$ denote a dual system over R , and suppose that \mathcal{X} is partially ordered by a pointed convex cone K . The *dual cone* K^+ and its *quasi-interior* K^{+i} are defined as

$$K^+ = \{f \in \mathcal{F} : \langle k, f \rangle \geq 0 \text{ for all } k \in K\}$$

and

$$K^{+i} = \{f \in \mathcal{F} : \langle k, f \rangle > 0 \text{ for all } k \in K \setminus \{\theta\}\}.$$

If the set K^{+i} is nonempty, we say that K admits strictly positive linear functionals. Throughout this work, it will be assumed that K^{+i} is nonempty. For general conditions under which such an assumption is valid, see [5, Prop. 2.7], [14, p. 58], [16, Thm. 2.7], [18, Thm. 2.8], and [22, pp. 26–27]. We mention specifically that the positive cones in the normed vector lattices R^n , $C[a, b]$, ℓ^p , and L^p all admit strictly positive linear functionals.

The following notions are fundamental in the study of vector optimization. Definition 2.1 is standard, and the terminology in Definition 2.2 for points that can be supported by strictly positive linear functionals was introduced in [9].

DEFINITION 2.1. Let \mathcal{X} be a vector space partially ordered by a pointed convex cone K , and let A be a nonempty subset of \mathcal{X} . A point $a \in A$ is said to be an *efficient* (*maximal*) *point* of A if $A \cap (\{a\} + K) = \{a\}$.

DEFINITION 2.2. Let $\langle \mathcal{X}, \mathcal{F} \rangle$ be a dual system over R , let A be a nonempty subset of \mathcal{X} , and suppose that \mathcal{X} is partially ordered by a pointed convex cone K . A point $a_0 \in A$ is said to be a *positive proper efficient point* of A if there exists $f \in K^{+i}$ such that $\langle a_0, f \rangle \geq \langle a, f \rangle$ for all $a \in A$.

We denote the set of efficient points of A by $E(A)$ and the set of positive proper efficient points of A by $\text{Pos}(A)$. In Lemma 2.3 some elementary observations regarding the sets $E(A)$ and $\text{Pos}(A)$ are given. The proof is left to the reader.

LEMMA 2.3. Let $\langle \mathcal{X}, \mathcal{F} \rangle$ be a dual system over R , and suppose that \mathcal{X} is partially ordered by a pointed convex cone K .

- (a) If $A \subseteq \mathcal{X}$, then $\text{Pos}(A) \subseteq E(A)$.
- (b) If A and C are nonempty subsets of \mathcal{X} satisfying $A \subseteq C \subseteq A - K$, then $E(A) = E(C)$ and $\text{Pos}(A) = \text{Pos}(C)$.
- (c) Suppose \mathcal{X} is equipped with a topology such that $\langle \cdot, f \rangle$ is continuous for all $f \in \mathcal{F}$, and suppose that K^{+i} is nonempty. If C is a compact subset of \mathcal{X} , then $\text{Pos}(C)$ is nonempty.

We next introduce the notion of a D-cone, the properties of which are fundamental to the proof of the density theorem (Theorem 2.6). The letter "D" in the name emphasizes the role that property (b) of the definition plays in the proof of the theorem; namely, to ensure that a set of saddle points which is constructed is indexed by a directed set ("D" for directed) so as to form a net.

If f and g are in \mathcal{F} , we use the notation $f \leq g$ to indicate that $g - f \in K^+$.

DEFINITION 2.4. Let $\langle \mathcal{X}, \mathcal{F} \rangle$ be a dual system over R , and let \mathcal{F} be equipped with the topology induced by \mathcal{X} . A pointed convex cone K in \mathcal{X} is said to be a *D-cone* if there exists a nonempty subset D of K^{+i} satisfying the following three conditions:

- (a) the set D is contained in a compact convex subset of \mathcal{F} ;
- (b) for every $f, g \in D$, there exists $h \in D$ such that $h \leq f$ and $h \leq g$; and
- (c) if $\langle x, f \rangle \geq 0$ for all $f \in D$, then $x \in K$.

We remark that even if the assumptions that K is pointed and convex are not explicitly mentioned in the above definition, they are implicitly implied by other assumptions. Indeed, the fact that K^{+i} is nonempty implies that K is pointed; and condition (c), together with the fact that D is a subset of K^{+i} , implies that

$$K = \bigcap_{f \in D} \{x \in \mathcal{X} : \langle x, f \rangle \geq 0\}.$$

Hence, K is convex. Moreover, if \mathcal{X} is equipped with a topology such that $\langle \cdot, f \rangle$ is continuous for all $f \in \mathcal{F}$, then K is closed.

In Lemma 2.5 sufficient conditions for a cone to be a D-cone are given. An important consequence of the lemma is that in a normed space any closed convex cone that admits strictly positive continuous linear functionals is a D-cone. Thus, in particular, the nonnegative orthants in R^n , $C[a, b]$, ℓ^p , and L^p , $1 \leq p \leq \infty$, are D-cones. It is also interesting to note that in addition to the above-mentioned normed spaces, the lemma also implies that the nonnegative orthants in the non-locally convex spaces ℓ^p , $0 < p < 1$ (e.g., see [25, p. 82]) are D-cones.

LEMMA 2.5. *Let \mathcal{X} be a locally bounded Hausdorff topological vector space such that \mathcal{X}^* separates points of \mathcal{X} , and let K be a weakly-closed convex cone in \mathcal{X} such that the set*

$$K^{+i} = \{f \in \mathcal{X}^* : f(k) > 0 \text{ for all } k \in K \setminus \{\theta\}\}$$

is nonempty. Then K is a D-cone.

Proof. Let V be a bounded neighborhood of the origin in \mathcal{X} . By the Banach-Alaoglu Theorem (e.g., see [25, Thm. 3.15]) the polar

$$V^\circ = \{f \in \mathcal{X}^* : |f(v)| \leq 1 \text{ for all } v \in V\}$$

of V is weak-star compact. Also, since V is bounded, it follows that $\sup\{|f(v)| : v \in V\}$ is finite for all $f \in \mathcal{X}^*$. Thus $K^{+i} \cap V^\circ$ is nonempty. Now, let $p \in K^{+i} \cap V^\circ$, and define

$$D = \bigcup_{n=1}^{\infty} (\{(1/n)p\} + K^+) \cap V^\circ.$$

Then D is a nonempty subset of K^{+i} contained in the weak-star compact set V° . It must be shown that conditions (b) and (c) of Definition 2.4 are satisfied.

Suppose $f, g \in D$. Then $f = (1/r)p + k_1^+$ and $g = (1/s)p + k_2^+$ for some positive integers r and s and some $k_1^+, k_2^+ \in K^+$. Choose $t = \max\{r, s\}$ and define $h = (1/t)p$. Then $h \in D$, $h \leq f$ and $h \leq g$. Hence, condition (b) is satisfied.

Suppose $x \notin K$. Since \mathcal{X}_w (\mathcal{X} with the weak topology) is a locally convex space, and since K is weakly closed and convex, there exists $q \in (\mathcal{X}_w)^*$ such that $q(x) < 0 \leq q(k)$ for all $k \in K$. The fact that \mathcal{X}^* separates points of \mathcal{X} implies that $\mathcal{X}^* = (\mathcal{X}_w)^*$ (e.g., see [25, pp. 62–63]) and hence, $q \in K^+$. Now, since V is bounded, we may assume that $\sup\{|q(v)| : v \in V\} < 1/2$. Then for n sufficiently large, we have $(1/n)p + q \in D$ and $((1/n)p + q)(x) < 0$. Hence, condition (c) is satisfied. \square

We now state and prove the main theorem of this section. The proof of the theorem uses essentially the same argument as the proofs of the density theorems given by Radner [24, p. 352], Borwein [3, Thm. 2], and Dauer and Gallagher [9, Thm. 4.5], which are all in the setting of normed spaces. (Radner's result is specifically stated for ℓ^∞ .) Proving the theorem in the more abstract setting of a dual system emphasizes that the essence of the argument is independent of any norm structure, and it clarifies the precise properties that the ordering cone must satisfy in order for the argument to remain valid; in particular, the cone must be a D-cone.

The notation $\text{cl}[S]$ is used to denote the closure of a subset S of a topological space.

THEOREM 2.6. *Let $\langle \mathcal{X}, \mathcal{F} \rangle$ be a dual system over R , let \mathcal{F} be equipped with the topology induced by \mathcal{X} , let \mathcal{X} be equipped with a topology such that $\langle \cdot, f \rangle$ is continuous for all $f \in \mathcal{F}$, and let \mathcal{X} be partially ordered by a D-cone K . If A is a subset of \mathcal{X} , and if there exists a compact convex subset C of \mathcal{X} such that $A \subseteq C \subseteq A - K$, then $E(A) \subseteq \text{cl}[\text{Pos}(A)]$.*

The following lemma will be used in the proof of Theorem 2.6; it is a corollary of a minimax theorem due to Fan [11, Thm. 1].

LEMMA 2.7. *Let A and B be compact convex sets, each in a topological vector space, and let Φ be a real-valued function defined on $A \times B$. Suppose that for each $b \in B$, $\Phi(a, b)$ is a continuous convex function on A , and for each $a \in A$, $\Phi(a, b)$ is a continuous concave function on B . Then there exists a pair $(a_0, b_0) \in A \times B$ satisfying*

$$\Phi(a_0, b) \leq \Phi(a_0, b_0) \leq \Phi(a, b_0)$$

for all $a \in A$ and all $b \in B$.

Proof of Theorem 2.6. Without loss of generality it will be shown that if $\theta \in E(A)$, then $\theta \in \text{cl}[\text{Pos}(A)]$. Since K is a D-cone, there exists a nonempty subset D of K^{+i} satisfying (a), (b), and (c) in Definition 2.4. Let B be the compact convex subset containing D , and for each $p \in D$ define

$$B(p) = \{f \in B : f \geq p\}.$$

Then $B(p)$ is a compact convex subset of \mathcal{F} . Since C is also compact and convex, Lemma 2.7 implies that there exists $c_p \in C$ and $f_p \in B(p)$ satisfying

$$(2.3) \quad \langle c, f_p \rangle \leq \langle c_p, f_p \rangle \leq \langle c_p, f \rangle$$

for all $c \in C$ and all $f \in B(p)$. Since $f_p \in K^{+i}$, it follows that $c_p \in \text{Pos}(C)$. Also, since $A \subseteq C \subseteq A - K$, it follows by Lemma 2.3 that $c_p \in \text{Pos}(A)$.

Since the set D satisfies condition (b) in Definition 2.4, the pair (D, \leq) is a directed set (e.g., see [7, p. 377]). Hence, the set $\{c_p : p \in D\}$ is a net in $\text{Pos}(A)$. Since C is compact, the net has a cluster point, say $\bar{c} \in C$. It follows that $\bar{c} \in \text{cl}[\text{Pos}(A)]$.

To finish the proof, we show that $\bar{c} = \theta$. Since $\theta \in E(A) = E(C)$, it suffices to show that $\bar{c} \in K$. Thus, by (c) in Definition 2.4, we need only show that $\langle \bar{c}, g \rangle \geq 0$ for all $g \in D$. To this end, let $g \in D$ and $\varepsilon > 0$ be given.

Since \bar{c} is a cluster point of $\{c_p : p \in D\}$ and since $\langle \cdot, g \rangle$ is continuous, there exists $r \in D$, $r \leq g$ such that

$$\langle \bar{c}, g \rangle > \langle c_r, g \rangle - \varepsilon.$$

Since $r \leq g$, we have $g \in B(r)$. Hence, since $\theta \in C$, inequality (2.3) implies that $\langle c_r, g \rangle \geq 0$. Thus, $\langle \bar{c}, g \rangle > -\varepsilon$. Since ε and g are arbitrary, it follows that $\langle \bar{c}, g \rangle \geq 0$ for all $g \in D$. Hence, $\bar{c} = \theta$. \square

As an immediate corollary of Theorem 2.6 and Lemma 2.5, we obtain the following generalization of a result of Borwein [4, Thm. 5]. As was mentioned in the Introduction, it seems that Borwein's result is not widely known.

COROLLARY 2.8. *Let \mathcal{X} be a locally bounded Hausdorff topological vector space such that \mathcal{X}^* separates points of \mathcal{X} , let K be a weakly closed convex cone in \mathcal{X} such that the set K^{+i} is nonempty, and let A be a subset of \mathcal{X} . If there exists a compact (respectively, weakly compact) convex subset C of \mathcal{X} such that $A \subseteq C \subseteq A - K$, then $E(A) \subseteq \text{cl}[\text{Pos}(A)]$ (respectively, $E(A) \subseteq \text{weak-cl}[\text{Pos}(A)]$).*

In Corollary 2.10 a local version of Corollary 2.8 is given for the case when \mathcal{X} is normed. The following elementary lemma is used in its proof.

LEMMA 2.9. *Let \mathcal{X} be a normed space and let A be a convex subset of \mathcal{X} . Let $\varepsilon > 0$ and $\bar{a} \in A$ be given, and define*

$$A(\varepsilon) = \{a \in A : \|\bar{a} - a\| < \varepsilon\}.$$

Let $f : A \rightarrow \mathbb{R}$ be a concave function, and suppose that there exists $a^ \in A(\varepsilon)$ satisfying*

$$f(a^*) \geq f(a) \quad \text{for all } a \in A(\varepsilon).$$

Then

$$f(a^*) \geq f(a) \quad \text{for all } a \in A.$$

Proof. If $a \in A$, there exists t , $0 \leq t < 1$, such that $ta^* + (1-t)a \in A(\varepsilon)$. It follows that $f(a^*) \geq f(a)$. \square

COROLLARY 2.10. *Let \mathcal{X} be a real normed space partially ordered by a closed convex cone K such that K^{+i} is nonempty, let A be a nonempty subset of \mathcal{X} , and let $\bar{a} \in A$ be an efficient point of A . If either*

- (a) *A is convex and the set $\bar{A}(t) = \{a \in A : \|a - \bar{a}\| \leq t\}$ is compact for some $t > 0$, or*
- (b) *$A - K$ is convex and the set $(\overline{A - K})(t) = \{x \in A - K : \|x - \bar{a}\| \leq t\}$ is compact for some $t > 0$,*

then $\bar{a} \in \text{cl}[\text{Pos}(A)]$.

Proof. (a) Let ε , $0 < \varepsilon \leq t$, be given; and define

$$A(\varepsilon) = \{a \in A : \|\bar{a} - a\| < \varepsilon\}.$$

Since $\bar{A}(t)$ is compact and convex, and since \bar{a} is an efficient point of $\bar{A}(t)$, Corollary 2.8 implies that $\bar{a} \in \text{cl}[\text{Pos}(\bar{A}(t))]$. Hence, there exists $a^* \in A(\varepsilon)$ and $f \in K^{+i}$ such that $f(a^*) \geq f(a)$ for all $a \in A(\varepsilon)$. By Lemma 2.9, it follows that $f(a^*) \geq f(a)$ for all $a \in A$; that is, $a^* \in \text{Pos}(A)$. Since ε is arbitrary, it follows that $\bar{a} \in \text{cl}[\text{Pos}(A)]$.

(b) This follows directly from part (a) and the set equalities $E(A) = E(A - K)$ and $\text{Pos}(A) = \text{Pos}(A - K)$. \square

3. Density for a space ordered by a cone with a bounded base. Throughout this section, \mathcal{X} will denote a (Hausdorff) locally convex topological vector space and K will denote a cone in \mathcal{X} . We say that a subset B of K is a *base* for K if B is convex, $\theta \notin \text{cl}[B]$, and

$$K = \text{cone}(B) \stackrel{\text{def}}{=} \{\lambda b : \lambda \geq 0, b \in B\}.$$

It is easily established that in a locally convex space, a based cone is convex and admits strictly positive continuous linear functionals (e.g., see [5, Prop. 2.7]). If, in addition, the base is closed and bounded, then K is closed [15, p. 121]. Hence, by Lemma 2.5, in a normed space, a cone with a closed bounded base is necessarily a D-cone. But in general, a D-cone need not have a bounded base. For example, the nonnegative orthants in $C[a, b]$, ℓ^p , and L^p , for $1 < p \leq \infty$, are D-cones but do not have bounded bases (e.g., see [9, Cors. 3.4 and 3.5] and [15, pp. 121–122 and 169]).

In this section we obtain density results similar to those given in the previous section, but are able to relax the hypotheses on the set A by assuming that the ordering cone has a closed bounded base. The primary examples of partially ordered locally convex spaces in which the ordering cones have closed bounded bases are R^n , ℓ^1 , and L^1 with their natural orderings and with either their norm or weak topologies. Other examples of closed and bounded based cones arise naturally in certain dual spaces. Indeed, if K is a cone in a locally convex space \mathcal{X} such that the interior of K is nonempty, then the dual cone $K^+ = \{f \in \mathcal{X}^* : f(k) \geq 0 \text{ for all } k \in K\}$ has a weak-star closed and bounded base (i.e., a weak-star compact base) [15, p. 23]. Thus, in particular, if K denotes the nonnegative orthant in $C[a, b]$, ℓ^∞ , or L^∞ , then the dual cone K^+ in the associated dual space has a weak-star closed bounded base. There is, of course, no shortage of closed and bounded based cones. Indeed, one can construct such cones from any closed, bounded convex set B not containing θ ; simply let $K = \text{cone}(B)$. For a thorough study of cones with bounded bases, see Jameson [15, pp. 120–126].

The main result of this section is Theorem 3.1, which is in the setting of a locally convex space. This result was recently obtained by Petschke [23] for normed spaces. It is important to note that the proof given here is substantially different from that given by Petschke. Indeed, Petschke shows that in a normed space $(\mathcal{X}, \|\cdot\|)$ a cone K with a closed bounded base is representable as a Bishop–Phelps cone; that is, there exists $f \in \mathcal{X}^*$ and a norm $p(\cdot)$ on \mathcal{X} which is equivalent to $\|\cdot\|$ such that

$$K = \{x \in \mathcal{X} : p(x) \leq f(x)\}.$$

Once this representation is established, the density result given in Jahn [13, Thm. 3.1], which requires the cone to be a Bishop–Phelps cone, is utilized. For completeness, we mention that the proof given by Jahn for his result involves an argument using weakly lower semicontinuous convex functions and subgradients to generate positive proper efficient points arbitrarily close to a prespecified efficient point. In contrast, the proof of Theorem 3.1 given below relies on standard separation theorems to generate such points.

THEOREM 3.1. *Let \mathcal{X} be a locally convex topological vector space partially ordered by a cone K with a closed bounded base, and let A be a subset of \mathcal{X} . If there exists a weakly compact convex subset C of \mathcal{X} such that $A \subseteq C \subseteq A - K$, then $E(A) \subseteq \text{cl}[\text{Pos}(A)]$.*

Proof. Without loss of generality, assume $\theta \in E(A)$. Let W be a convex open neighborhood of θ . We show that $W \cap \text{Pos}(A) \neq \emptyset$. Since \mathcal{X} is locally convex, there

exist convex open neighborhoods U and V of θ such that $\text{cl}[V] \subseteq W$ and $U + U \subseteq V$ (e.g., see [25, pp. 9–10]).

Let B denote a closed bounded base for K . Since B is bounded, we may assume that $B \subseteq U$. Since $\theta \in E(A) = E(C)$, we have $C \cap K = \{\theta\}$. Hence, $C \cap B = \emptyset$; and consequently, letting $\overline{C} := C \cap \text{cl}[V]$, we have $\overline{C} \cap B = \emptyset$. But \overline{C} is weakly compact and convex, and B is weakly closed and convex; so there exists $g \in \mathcal{X}^*$ and $\bar{c} \in \overline{C}$ such that

$$(3.1) \quad g(\bar{c}) = \max\{g(c) : c \in \overline{C}\} < \inf\{g(b) : b \in B\}$$

(e.g., see [25, Thm. 3.4(b)]).

To finish the proof, it suffices to show that $\bar{c} \in \text{Pos}(A)$. From (3.1), there exists a convex open neighborhood N of θ such that $N \subseteq U$ and

$$\widehat{B} := B + N \subseteq \{x \in \mathcal{X} : g(x) > g(\bar{c})\}.$$

Let $P = \text{cone}(\widehat{B})$. Then $K \setminus \{\theta\} \subseteq \text{int}[P]$ (the interior of P). We next show that $P \cap (C - \{\bar{c}\}) = \{\theta\}$. For the sake of contradiction, suppose there exists $\lambda > 0$ and $c \in C \setminus \{\bar{c}\}$ such that $c - \bar{c} = \lambda \hat{b}$ for some $\hat{b} \in \widehat{B}$. Since C is convex and since $\theta \in C$, the vector $z := (1 + \lambda)^{-1}c \in C$. Also $z \in \text{cl}[V]$ and $g(z) > g(\bar{c})$. But this contradicts (3.1); hence, $P \cap (C - \{\bar{c}\}) = \{\theta\}$. Consequently, $\text{int}[P] \cap (C - \{\bar{c}\}) = \emptyset$. Thus, there exists $f \in \mathcal{X}^*$, $f \neq 0$, such that

$$(3.2) \quad \sup\{f(x) : x \in (C - \{\bar{c}\})\} \leq 0 \leq \inf\{f(p) : p \in P\}.$$

Moreover, $f(p) > 0$ for all $p \in \text{int}[P]$ (e.g., see [25, Thm. 3.4(a)]). Since $K \setminus \{\theta\} \subseteq \text{int}[P]$, $f(k) > 0$ for all $k \in K \setminus \{\theta\}$. Also, it follows from (3.2) that $f(\bar{c}) \geq f(c)$ for all $c \in C$. Hence, $\bar{c} \in \text{Pos}(C) = \text{Pos}(A)$. \square

The following corollary provides a local version of Theorem 3.1 in a normed space setting. Its proof is nearly identical to the proof of Corollary 2.10 and is therefore omitted.

COROLLARY 3.2. *Let \mathcal{X} be a real normed space partially ordered by a cone K with a closed bounded base, let A be a nonempty subset of \mathcal{X} , and let $\bar{a} \in A$ be an efficient point of A . If either*

- (a) *A is convex and the set $\overline{A}(t) = \{a \in A : \|a - \bar{a}\| \leq t\}$ is weakly compact for some $t > 0$, or*
- (b) *$A - K$ is convex and the set $\overline{(A - K)}(t) = \{x \in A - K : \|x - \bar{a}\| \leq t\}$ is weakly compact for some $t > 0$,*

then $\bar{a} \in \text{cl}[\text{Pos}(A)]$.

The next corollary, in the setting of a reflexive normed space, generalizes the finite dimensional results of Hartley [12, Thm. 5.5] and Bitran and Magnanti [2, Cor. 3.1]. (Also, compare with [13, Cor. 3.4].)

COROLLARY 3.3. *Let \mathcal{X} be a reflexive Banach space partially ordered by a cone K with a closed bounded base, and let A be a subset of \mathcal{X} such that either A is closed and convex or $A - K$ is closed and convex. Then $\text{Pos}(A)$ is norm dense in $E(A)$.*

We conclude this section by mentioning that Theorem 3.1 can be applied to prove an important result of Klee, which states that the support points of a locally weakly compact convex subset A of a locally convex space are dense in the boundary of A [17, Thm. 9.11]. The authors will present this proof in a forthcoming paper.

4. Conclusion. As was mentioned in the introduction and as is revealed when comparing Corollary 2.8 with Theorem 3.1, if A is a subset of a normed space, then existing results giving sufficient conditions for the set $\text{Pos}(A)$ to be norm dense in the set $E(A)$ either require strong assumptions on the ordering cone and relatively weak assumptions on the set A , or relatively weak assumptions on the ordering cone and strong assumptions on the set A . It is not known whether such trade-offs are absolutely necessary. From the results presented here, an obvious question that arises is the following: Is it possible to show that $\text{Pos}(A)$ is norm dense in $E(A)$ if A is only assumed to be weakly compact and convex (as in Theorem 3.1) and the ordering cone is only assumed to be closed and convex and to admit strictly positive continuous linear functionals (as in Corollary 2.8)? It would seem that the method of proof or counterexample needed to answer this question would have important implications with regard to other functional analytic questions as well.

REFERENCES

- [1] K. J. ARROW, E. W. BARANKIN, AND D. BLACKWELL, *Admissible points of convex sets*, in Contributions to the Theory of Games, Vol. II (H. W. Kuhn and A. W. Tucker, eds.), Princeton University Press, Princeton, NJ, 1953, pp. 87–92.
- [2] G. R. BITRAN AND T. L. MAGNANTI, *The structure of admissible points with respect to cone dominance*, J. Optim. Theory Appl., 29 (1979), pp. 573–614.
- [3] J. M. BORWEIN, *The geometry of Pareto efficiency over cones*, Math. Oper. Statist., Ser. Optim., 11 (1980), pp. 235–248.
- [4] ———, *On the existence of Pareto efficient points*, Math. Oper. Res., 8 (1983), pp. 64–73.
- [5] ———, *Continuity and differentiability properties of convex operators*, Proc. London Math. Soc., 44 (1982), pp. 420–444.
- [6] G. CHICHILNISKY AND P. J. KALMAN, *Applications of functional analysis to models of efficient allocation of economic resources*, J. Optim. Theory Appl., 30 (1980), pp. 19–32.
- [7] J. B. CONWAY, *A Course in Functional Analysis*, Springer-Verlag, New York, 1985.
- [8] N. O. DACUNHA AND E. POLAK, *Constrained minimization under vector-valued criteria in finite dimensional spaces*, J. Math. Anal. Appl., 19 (1967), pp. 103–124.
- [9] J. P. DAUER AND R. J. GALLAGHER, *Positive proper efficient points and related cone results in vector optimization theory*, SIAM J. Control Optim., 28 (1990), pp. 158–172.
- [10] R. DURIER, *Weighting factor results in vector optimization*, J. Optim. Theory Appl., 58 (1988), pp. 411–430.
- [11] K. FAN, *Minimax theorems*, Proc. Nat. Acad. Sci., U.S.A., 39 (1953), pp. 42–47.
- [12] R. HARTLEY, *On cone-efficiency, cone-convexity and cone-compactness*, SIAM J. Appl. Math., 34 (1978), pp. 211–222.
- [13] J. JAHN, *A generalization of a theorem of Arrow, Barankin, and Blackwell*, SIAM J. Control Optim., 26 (1988), pp. 999–1005.
- [14] ———, *Mathematical Vector Optimization*, Peter Lang, Frankfurt, Germany, 1986.
- [15] G. JAMESON, *Ordered Linear Spaces*, Springer-Verlag, New York, 1970.
- [16] V. L. KLEE, *Separation properties of convex cones*, Proc. Amer. Math. Soc., 6 (1955), pp. 313–318.
- [17] ———, *Convex sets in linear spaces*, Duke Math. J., 18 (1951), pp. 443–466.
- [18] M. G. KREIN AND M. A. RUTMAN, *Linear operators leaving invariant a cone in a Banach space*, Translation Number 26, American Mathematical Society, 1950.
- [19] M. MAJUMDAR, *Some general theorems on efficiency prices with an infinite dimensional commodity space*, J. Econom. Theory, 5 (1972), pp. 1–13.
- [20] B. PELEG, *Topological properties of the efficient point set*, Proc. Amer. Math. Soc., 35 (1972), pp. 531–536.
- [21] ———, *Efficiency prices for optimal consumption plans*, II, Israel J. Math., 9 (1971), pp. 222–234.
- [22] A. L. PERESSINI, *Ordered Topological Vector Spaces*, Harper and Row, New York, 1967.
- [23] M. PETSCHKE, *On a theorem of Arrow, Barankin, and Blackwell*, SIAM J. Control Optim., 28 (1990), pp. 395–401.

- [24] R. RADNER, *A note on maximal points of convex sets in ℓ^∞* , in Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, CA, 1967, pp. 351–354.
- [25] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [26] W. SALZ, *Eine topologische Eigenschaft der effizienten Punkte konvexer Mengen*, Operat. Res. Verfahren, XXIII (1976), pp. 197–202.
- [27] H. H. SCHAEFER, *Topological Vector Spaces*, Springer-Verlag, New York, 1986.

LOWER SEMICONTINUOUS SOLUTIONS OF HAMILTON–JACOBI–BELLMAN EQUATIONS*

HÉLÈNE FRANKOWSKA†

Abstract. The value function of Mayer’s problem arising in optimal control is investigated, and lower semicontinuous solutions of the associated Hamilton–Jacobi–Bellman equation are defined in three (equivalent) ways. Under quite weak assumptions about the control system, the value function is the unique solution. Moreover, it is stable with respect to perturbations of the control system and the cost. It coincides with the viscosity solution whenever it is continuous.

Key words. Hamilton–Jacobi equation, Mayer’s problem, viability theory, viscosity solution

AMS(MOS) subject classifications. 35B37, 49C20, 49J15, 49J45, 90C39

1. Introduction. Consider the following Hamilton–Jacobi–Bellman (HJB) equation:

$$(1) \quad -\frac{\partial V}{\partial t}(t, x) + H\left(t, x, -\frac{\partial V}{\partial x}(t, x)\right) = 0, \quad V(T, \cdot) = g(\cdot)$$

associated with the Mayer problem arising in control theory:

$$\text{minimize } \{ g(x(T)) \mid x(T) \in K \}$$

over all solutions of the control system

$$(2) \quad x' = f(t, x, u(t)), \quad u(t) \in U$$

satisfying the initial condition

$$(3) \quad x(0) = \xi_0,$$

where K is a given set (called target).

By a simple change of variables, the classical Bolza problem

$$\text{minimize } \left\{ g(x(T)) + \int_0^T L(t, x(t), u(t)) dt \mid x(T) \in K \right\}$$

over all state-control solutions (x, u) of (2), (3) may be reduced to the Mayer problem.

In (1) the Hamiltonian H is given by

$$H(t, x, p) = \sup_{u \in U} \langle p, f(t, x, u) \rangle.$$

The value function for Mayer’s problem is defined by

$$(4) \quad V(t_0, x_0) = \inf \{ g(x(T)) \mid x \text{ is a solution of (2), } x(t_0) = x_0, x(T) \in K \}$$

In general, V is merely lower semicontinuous and is equal to $+\infty$ at all points from which it is impossible to reach the target K . In fact, we can even avoid explicitly mentioning the target K in (4) by setting $g(x) = +\infty$ whenever $x \notin K$.

* Received by the editors February 25, 1991; accepted for publication (in revised form) September 17, 1991.

† Centre de Recherche de Mathématiques de la Décision, Université de Paris-Dauphine, 75775 Paris Cx 16, France.

The value function allows us to single out optimal trajectories since it is nondecreasing along solutions of (2) and is constant along optimal solutions. These two properties and the final value $V(T, \cdot) = g(\cdot)$ characterize the value function.

When V is differentiable, it solves the partial differential equation (1). When it is merely lower semicontinuous, the solution of the HJB equation must be defined in such way that under quite general assumptions on H and g , V is the unique solution of (1).

In [12], [14] the notion of viscosity solution was introduced to deal with continuous solutions of the Hamilton–Jacobi equation. The basic idea consists in replacing the gradient by *superdifferential* and *subdifferential* and the equality in (1) by two inequalities. In [13] several results concerning uniqueness of solutions in the class of uniformly continuous functions is given.

A different approach, based on Dini derivatives, was developed in [25], [26] for Lipschitz solutions.

Since the value function is merely lower semicontinuous, an extension of viscosity solutions to semicontinuous functions is needed.

In [16]–[19] a study using contingent derivatives (which are extensions of Dini derivatives to semicontinuous functions and are defined in §2) was proposed. It was shown that a function V satisfying two contingent inequalities (called contingent solution of the HJB equation) has properties of the value function: it is nondecreasing along solutions of (2) and is constant along at least one solution. It is unique whenever the final value $V(T, \cdot)$ is fixed. Proofs are based on viability theory ([2]–[4]). Relationships with viscosity solutions were also discussed. Contingent derivatives have further applications in control theory. For instance, in some cases the optimal synthesis may be obtained using the directional derivatives of the value function (see [19], [10]).

Results obtained in [16]–[19] do not allow us to deal with arbitrary lower semicontinuous functions. Namely, one of the contingent inequalities

$$\forall (t, x) \in \text{Dom}(V), \quad t < T, \quad \sup_{u \in U} D_{\uparrow}(-V)(t, x)(1, f(t, x, u)) \leq 0,$$

which implies an invariance property of the epigraph of $-V$, allows us to deal only with upper semicontinuous V .

In this paper we replace it by a new inequality, specifically,

$$\forall (t, x) \in \text{Dom}(V), \quad t > 0, \quad \sup_{u \in U} D_{\uparrow}V(t, x)(-1, -f(t, x, u)) \leq 0,$$

which yields an invariance property of the epigraph of V . Such a modification leads to much stronger results. On one hand, it allows us to provide a characterization of the lower semicontinuous value function in terms of contingent derivatives and the final value $V(T, \cdot) = g(\cdot)$ (see §2 for the main result); on the other hand, every continuous contingent solution of the HJB equation turns out to be its viscosity solution (this is shown in §7). In particular, it implies uniqueness results for viscosity solutions of HJB equations with convex Hamiltonian $H(t, x, \cdot)$ under less restrictive assumptions than in [13].

Recently, in [8], [9], a modification of the concept of viscosity solutions for semicontinuous functions was proposed. The approach is based on a construction of “touching from one side” functions, which is usual for viscosity solutions theory. Such a modified definition of a solution states that a lower semicontinuous function V is a solution of (1) if for all $(p_t, p_x) \in \mathbf{R} \times \mathbf{R}^n$ in the subdifferential of V at $(t, x) \in]0, T[\times \mathbf{R}^n$,

$$-p_t + H(t, x, -p_x) = 0$$

and for all $\bar{x} \in \mathbf{R}^n$,

$$(5) \quad V(0, \bar{x}) = \liminf_{(t,x) \rightarrow (0+, \bar{x})} V(t, x) \ \& \ g(\bar{x}) = \liminf_{(t,x) \rightarrow (T-, \bar{x})} V(t, x).$$

We have uniqueness of lower semicontinuous solutions of (1), when H is, loosely speaking, Lipschitz and convex with respect to p . Furthermore, a continuous function V is a viscosity solution if and only if it satisfies the above properties. This result is improved in this paper by relaxing assumptions about the Hamiltonian H .

We discuss three equivalent definitions of lower semicontinuous solutions of HJB equation: the one from [8], a modified “contingent definition,” and the one with (5) replaced by the following: For all (p_t, p_x) in the subdifferential of V at $(0, \bar{x})$,

$$-p_t + H(0, x, -p_x) \geq 0,$$

and for all (p_t, p_x) in the subdifferential of V at (T, \bar{x}) ,

$$-p_t + H(T, x, -p_x) \leq 0.$$

As in [16], the modified “contingent definition” yields monotonicity. The approach we use is still based on viability theory. We describe the dynamics by differential inclusions since, by the well-known arguments, the control system (2) can be reduced to the differential inclusion

$$x' \in F(t, x) := f(t, x, U).$$

The outline of the paper is as follows: In §2 we state the main result. Section 3 is devoted to the monotone behavior of contingent solutions and §4 is devoted to solutions formulated in terms of subdifferentials. In §5 we address the definition of solution proposed in [8] using alternative boundary conditions. A comparison with continuous viscosity solutions is provided in §6. In §7 we associate with a Hamilton–Jacobi–Bellman equation (1) (with convex Hamiltonian) an optimal control problem whose value function is the only solution of the HJB equation. This leads to both existence and uniqueness results for (1). Finally, in §8 the stability of value function is investigated. In view of the results of §7, this allows us to study stability of solutions of the HJB equation under perturbations of H and g .

2. Main theorem. Consider $T > 0$, a set-valued map $F : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ and the associated differential inclusion

$$(6) \quad x'(t) \in F(t, x(t)) \text{ almost everywhere.}$$

We denote by $S_{[t_0, T]}(x_0)$ the set of absolutely continuous solutions of (6) defined on $[t_0, T]$ and satisfying the initial condition $x(t_0) = x_0$.

Let an extended function $g : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ and $\xi_0 \in \mathbf{R}^n$ be given. Consider the following minimization problem (called *Mayer’s problem*):

$$(7) \quad \min \{g(x(T)) \mid x \text{ is a solution to (6), } x(0) = \xi_0\}.$$

The value function $\mathcal{V} : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{\pm\infty\}$ associated with it is defined by the following: for all $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$,

$$(8) \quad \mathcal{V}(t_0, x_0) = \inf \{g(x(T)) \mid x \in S_{[t_0, T]}(x_0)\}.$$

Denote by $B_R(0)$ the closed ball of center zero and radius $R \geq 0$. We impose the following assumptions

- (a) F is continuous with nonempty convex compact images;
- (b) $\exists k \in L^1(0, T)$ such that for almost all $t \in [0, T]$, we have
- (9) $\forall x \in \mathbf{R}^n, \sup_{v \in F(t, x)} \|v\| \leq k(t)(1 + \|x\|);$
- (c) $\forall R > 0, \exists c_R \in L^1(0, T)$ such that for almost all $t \in [0, T]$
 $F(t, \cdot)$ is $c_R(t)$ -Lipschitz on $B_R(0)$.

PROPOSITION 2.1. *If (9) holds true, then \mathcal{V} is lower semicontinuous and*

$$(10) \quad \forall (t_0, x_0) \in [0, T] \times \mathbf{R}^n, \quad \mathcal{V}(t_0, x_0) = \min \{g(x(T)) \mid x \in S_{[t_0, T]}(x_0)\}.$$

Furthermore,

$$(11) \quad \forall \bar{x} \in \mathbf{R}^n, \quad g(\bar{x}) = \liminf_{t \rightarrow T-, x \rightarrow \bar{x}} \mathcal{V}(t, x), \quad \mathcal{V}(0, \bar{x}) = \liminf_{t \rightarrow 0+, x \rightarrow \bar{x}} \mathcal{V}(t, x).$$

Proof. The first statement results from the compactness of $S_{[t_0, T]}(x_0)$ in $\mathcal{C}(t_0, T; \mathbf{R}^n)$ (see, for instance, [4], [5, p. 273]). To prove (11), consider $\bar{x} \in \mathbf{R}^n$ and let $t_i \rightarrow T-, \bar{x}_i \rightarrow \bar{x}$ be such that

$$\lim_{i \rightarrow \infty} \mathcal{V}(t_i, \bar{x}_i) = \liminf_{t \rightarrow T-, x \rightarrow \bar{x}} \mathcal{V}(t, x).$$

Consider $y_i \in \mathbf{R}^n$ and $x_i \in S_{[t_i, T]}(y_i)$ such that $x_i(T) = \bar{x}$. Then, by (9), $y_i \rightarrow \bar{x}$. Since $\mathcal{V}(t_i, y_i) \leq g(x_i(T)) = g(\bar{x})$, we obtain

$$\lim_{i \rightarrow \infty} \mathcal{V}(t_i, \bar{x}_i) \leq \liminf_{i \rightarrow \infty} \mathcal{V}(t_i, y_i) \leq g(\bar{x}) = V(T, \bar{x})$$

and deduce the first equality in (11) using the lower semicontinuity of V .

Let $x \in S_{[0, T]}(\bar{x})$ be such that $\mathcal{V}(0, \bar{x}) = g(x(T))$. Then $\mathcal{V}(0, \bar{x}) \equiv \mathcal{V}(t, x(t))$ for all $t \in [0, T]$, and the result follows from the lower semicontinuity of \mathcal{V} . \square

DEFINITION 2.2. Consider an extended function $\varphi : \mathbf{R}^n \mapsto \mathbf{R} \cup \{\pm\infty\}$.

- (i) The domain of φ , $\text{Dom}(\varphi)$, is the set of all x_0 such that $\varphi(x_0) \neq \pm\infty$.
- (ii) The subdifferential of φ at $x_0 \in \text{Dom}(\varphi)$ is given by

$$\partial_- \varphi(x_0) = \left\{ p \in \mathbf{R}^n \mid \liminf_{x \rightarrow x_0} \frac{\varphi(x) - \varphi(x_0) - \langle p, x - x_0 \rangle}{\|x - x_0\|} \geq 0 \right\}$$

(iii) The contingent epiderivative of φ at $x_0 \in \text{Dom}(\varphi)$ in the direction $u \in \mathbf{R}^n$ is given by

$$D_1 \varphi(x_0)(u) = \liminf_{h \rightarrow 0+, u' \rightarrow u} \frac{\varphi(x_0 + hu') - \varphi(x_0)}{h}.$$

Remark. The subdifferential was used in [12], [14] to define viscosity supersolutions of Hamilton–Jacobi equations. The contingent epiderivative was used in [1], [4] to investigate stability. See also [5, Chap. 6] for further properties.

Assume that F has nonempty compact images and define the Hamiltonian $H : [0, T] \times \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}$ by

$$(12) \quad H(t, x, p) = \max_{v \in F(t, x)} \langle p, v \rangle.$$

Then $H(t, x, \cdot)$ is convex and positively homogeneous. Furthermore, if F is upper semicontinuous (respectively, lower semicontinuous), then so is H .

Consider a function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$. We may always assume that V is defined on $\mathbf{R} \times \mathbf{R}^n$ by setting $V(t, x) = +\infty$, whenever $t \notin [0, T]$. In the theorem below we use Definition 2.2 with such an extension of V .

THEOREM 2.3. *Assume (9) and let $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semicontinuous function.*

Then the following five statements are equivalent:

- (i) V is the value function, i.e., $V = \mathcal{V}$;
- (ii) $V(T, \cdot) = g(\cdot)$ and for all $(t_0, x_0) \in \text{Dom}(V)$ we have
 $\forall x \in S_{[t_0, T]}(x_0), \forall t \in [t_0, T], V(t, x(t)) \geq V(t_0, x_0)$
 $\exists \bar{x} \in S_{[t_0, T]}(x_0), \forall t \in [t_0, T], V(t, \bar{x}(t)) \leq V(t_0, x_0)$;
- (iii) $V(T, \cdot) = g(\cdot)$ and for all $(t, x) \in \text{Dom}(V)$,
 $0 \leq t < T \implies \inf_{v \in F(t, x)} D_{\uparrow} V(t, x)(1, v) \leq 0$
 $0 < t \leq T \implies \sup_{v \in F(t, x)} D_{\uparrow} V(t, x)(-1, -v) \leq 0$;
- (iv) $\forall (t, x) \in]0, T[\times \mathbf{R}^n, \forall (p_t, p_x) \in \partial_- V(t, x), -p_t + H(t, x, -p_x) = 0$
 $\forall (p_t, p_x) \in \partial_- V(0, x), -p_t + H(0, x, -p_x) \geq 0$
 $\forall (p_t, p_x) \in \partial_- V(T, x), -p_t + H(T, x, -p_x) \leq 0$ and $V(T, \cdot) = g(\cdot)$;
- (v) $\forall (t, x) \in]0, T[\times \mathbf{R}^n, \forall (p_t, p_x) \in \partial_- V(t, x), -p_t + H(t, x, -p_x) = 0$
 $\forall \bar{x} \in \mathbf{R}^n, V(0, \bar{x}) = \liminf_{t \rightarrow 0+, x \rightarrow \bar{x}} V(t, x)$
 $\forall \bar{x} \in \mathbf{R}^n, V(T, \bar{x}) = \liminf_{t \rightarrow T-, x \rightarrow \bar{x}} V(t, x)$, and $V(T, \cdot) = g(\cdot)$.

Finally, if $\text{Dom}(V)$ is closed and the restriction of V to its domain is continuous and if

$$(13) \quad ([0, T] \times \mathbf{R}^n) \cap \text{Dom}(V) \subset \overline{([0, T[\times \mathbf{R}^n) \cap \text{Dom}(V)}}$$

then the above statements are equivalent to the following:

V is a viscosity solution of the Hamilton–Jacobi equation

$$-\frac{\partial V}{\partial t}(t, x) + H\left(t, x, -\frac{\partial V}{\partial x}(t, x)\right) = 0, \quad V(T, \cdot) = g(\cdot)$$

on $\text{Dom}(V)$.

Remark. We recall the definition of viscosity solutions in §6.

The fact that (i) \iff (ii) is well known (see, for instance, [15, §3.3]). We prove all other equivalences in §§3–6.

3. Monotone behavior of contingent solutions. Theorems 3.2 and 3.3 proved below yield equivalence of (ii) and (iii) in Theorem 2.3. We need the following definition.

DEFINITION 3.1. Let $K \subset \mathbf{R}^n$ be a nonempty subset and $x_0 \in K$. The contingent cone to K at x_0 , $T_K(x_0)$, is defined by

$$v \in T_K(x_0) \iff \liminf_{h \rightarrow 0+} d\left(v, \frac{K - x_0}{h}\right) = 0.$$

This notion was introduced by Bouligand in the early 1930s. See, for instance, [5, Chap. 4] for many properties of tangent cones.

It was shown in [5, p. 226] that for $\varphi : \mathbf{R}^n \rightarrow \mathbf{R} \cup \{\pm\infty\}$ and $x_0 \in \text{Dom}(\varphi)$

$$(14) \quad \mathcal{E}p(D_{\uparrow}\varphi(x_0)) = T_{\mathcal{E}p(\varphi)}(x_0, \varphi(x_0)),$$

where $\mathcal{E}p$ denotes the epigraph. Consider $F : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$.

THEOREM 3.2. *Let $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semicontinuous function. Assume that F is upper semicontinuous, that $F(t, x)$ is nonempty convex and compact for all $(t, x) \in \text{Dom}(V)$, and that for some integrable function $k : [0, T] \mapsto \mathbf{R}_+$,*

$$\forall (t, x) \in \text{Dom}(V), \quad \sup_{v \in F(t, x)} \|v\| \leq k(t)(1 + \|x\|).$$

Then the following two statements are equivalent

- (i) $\forall (t, x) \in \text{Dom}(V)$ with $t < T$, $\inf_{v \in F(t, x)} D_{\uparrow}V(t, x)(1, v) \leq 0$;
- (ii) $\forall (t_0, x_0) \in [0, T] \times \mathbf{R}^n$, $\exists \bar{x} \in S_{[t_0, T]}(x_0)$, $\forall t \in [t_0, T]$, $V(t, \bar{x}(t)) \leq V(t_0, x_0)$.

Proof. Assume that (i) holds true and fix $(t_0, x_0) \in \text{Dom}(V)$. Define the upper semicontinuous set-valued map $\hat{F} : \mathbf{R}_+ \times \mathbf{R}^n \times \mathbf{R} \rightsquigarrow \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}$ by

$$\hat{F}(t, x, z) = \begin{cases} \{1\} \times F(t, x) \times \{0\} & \text{when } t < T \\ [0, 1] \times \overline{\text{co}}(F(T, x) \cup \{0\}) \times \{0\} & \text{when } t \geq T \end{cases}$$

and consider the viability problem

$$(15) \quad \begin{aligned} (t, x, z)' &\in \hat{F}(t, x, z), \\ (t, x, z)(t_0) &= (t_0, x_0, V(t_0, x_0)), \\ (t, x, z) &\in \mathcal{E}p(V), \end{aligned}$$

By (14), for all $(t, x, z) \in \mathcal{E}p(V)$ we have $\hat{F}(t, x, z) \cap T_{\mathcal{E}p(V)}(t, x, z) \neq \emptyset$.

The viability theorem [22] (see also [4, p. 180], [2], [3]) yields that problem (15) has a solution

$$[t_0, T] \ni t \mapsto (t, \bar{x}(t), z(t)) \in \mathcal{E}p(V).$$

Thus $V(t, \bar{x}(t)) \leq z(t) = V(t_0, x_0)$ for all $t \in [t_0, T]$ and (ii) follows.

Conversely, assume that (ii) is satisfied. Fix $(t_0, x_0) \in \text{Dom}(V)$ with $t_0 < T$ and let \bar{x} be as in (ii). Let $h_n \rightarrow 0+$ be such that $[\bar{x}(t_0 + h_n) - \bar{x}(t_0)]/h_n$ converge to some v . By the mean value theorem [4, p. 21], $v \in F(t_0, x_0)$, and by (ii),

$$D_{\uparrow}V(t_0, x_0)(1, v) \leq 0. \quad \square$$

THEOREM 3.3. *Let $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semicontinuous function. Assume that F verifies (9).*

Then the following two statements are equivalent:

- (i) $\forall (t, x) \in \text{Dom}(V)$ with $t > 0$, $\sup_{v \in F(t, x)} D_{\uparrow}V(t, x)(-1, -v) \leq 0$;
- (ii) $\forall (t_0, x_0) \in [0, T] \times \mathbf{R}^n$, $\forall x \in S_{[t_0, T]}(x_0)$ and for all $t \in [t_0, T]$, $V(t_0, x_0) \leq V(t, x(t))$.

Proof. Assume (i) and fix $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$, $x \in S_{[t_0, T]}(x_0)$. Since (i) does not involve T , it is sufficient to prove the inequality in (ii) for $t = T$ and it is enough to consider the case $V(T, x(T)) < \infty$.

Let U denote the closed unit ball in \mathbf{R}^n . From [5, p. 380] there exists a continuous function $f : [0, T] \times \mathbf{R}^n \times U \mapsto \mathbf{R}^n$ such that

$$\begin{aligned} \forall (t, x) \in [0, T] \times \mathbf{R}^n, \quad F(t, x) &= f(t, x, U), \\ \forall u \in U, \quad f(t, \cdot, u) &\text{ is } 5nc_R(t) - \text{Lipschitz on } B_R(0) \text{ for a.e. } t \in [0, T], \\ \forall u, v \in U, \quad \|f(t, x, u) - f(t, x, v)\| &\leq 5n(\sup_{y \in F(t, x)} \|y\|) \|u - v\|. \end{aligned}$$

By [5, p. 316] for a measurable $u : [t_0, T] \mapsto U$, $x'(t) = f(t, x(t), u(t))$ almost everywhere. Consider a sequence of continuous maps $u_k : [t_0, T] \mapsto U$ converging to u in $L^1(t_0, T; U)$ and let x_k denote the solution of

$$x'_k(t) = f(t, x_k(t), u_k(t)), \quad t \in [t_0, T], \quad x_k(T) = x(T).$$

The Gronwall lemma yields that $x_k(t_0)$ converge to x_0 . On the other hand, the map $t \mapsto (T - t, x_k(T - t), V(T, x(T)))$ is the only solution of

$$(16) \quad \begin{aligned} \gamma'(t) &= -1, \\ y'(t) &= -f(T - t, y(t), u_k(T - t)), \\ z'(t) &= 0, \\ \gamma(0) &= T, \quad y(0) = x(T), \quad z(0) = V(T, x(T)). \end{aligned}$$

By (14) and (i) we know that

$$\forall (\gamma, x, z) \in \mathcal{E}p(V), \quad (-1, -f(\gamma, x, u_k(\gamma)), 0) \in T_{\mathcal{E}p(V)}(\gamma, x, z).$$

The map $(t, x) \leadsto \{-f(T - t, x, u_k(T - t))\}$ being continuous, the viability theorem [22] yields that (16) has at least one solution

$$[0, T - t_0] \ni t \mapsto (\gamma(t), y(t), z(t)) \in \mathcal{E}p(V).$$

Consequently,

$$\forall 0 \leq t \leq T - t_0, \quad (T - t, x_k(T - t), V(T, x(T))) \in \mathcal{E}p(V);$$

therefore, for all $0 \leq t \leq T - t_0$, $V(T, x(T)) \geq V(T - t, x_k(T - t))$. In particular, $V(t_0, x_k(t_0)) \leq V(T, x(T))$. Taking the limit when $k \rightarrow \infty$ and using that V is lower semicontinuous, we deduce (ii) for $t = T$.

Conversely, assume that (ii) is verified. Let $(t_0, x_0) \in \text{Dom}(V)$ be such that $t_0 > 0$. Fix $v \in F(t_0, x_0)$. Filippov's theorem [4, p. 120] implies that for some $\bar{h} > 0$ there exist $y_0 \in \mathbf{R}^n$ and $y \in S_{[t_0 - \bar{h}, t_0]}(y_0)$ such that $y(t_0) = x_0$ and

$$\lim_{h \rightarrow 0+} \frac{y(t_0 - h) - x_0}{h} = -v.$$

On the other hand, by (ii), for all $h \in [0, \bar{h}]$, $V(t_0 - h, y(t_0 - h)) \leq V(t_0, x_0)$. Consequently $D_{\uparrow} V(t_0, x_0)(-1, -v) \leq 0$. Since $v \in F(t_0, x_0)$ is arbitrary, statement (i) follows. \square

4. Subgradient form of the HJB equation. The equivalence (iii) \iff (iv) of Theorem 2.3 follows from Lemmas 4.3 and 4.4 given in this section. From [19] (see also [5, pp. 249, 253]) we have the following result.

PROPOSITION 4.1. *Let $\varphi : \mathbf{R}^n \mapsto \mathbf{R} \cup \{\pm\infty\}$ and $x_0 \in \text{Dom}(\varphi)$. Then the following statements are equivalent:*

- (i) $p \in \partial_- \varphi(x_0)$;
- (ii) $\forall u \in \mathbf{R}^n, \langle p, u \rangle \leq D_{\uparrow} \varphi(x_0)(u)$;
- (iii) $(p, -1) \in [T_{\mathcal{E}_p(\varphi)}(x_0, \varphi(x_0))]^-$ (the negative polar cone).

The following result can be deduced from [24, Theorem 1] (see also [23]).

LEMMA 4.2. *Consider an extended lower semicontinuous function $\varphi : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ and $x_0 \in \text{Dom}(\varphi)$. Let $(p, 0) \in [T_{\mathcal{E}_p(\varphi)}(x_0, \varphi(x_0))]^-$ be such that $p \neq 0$. Then for every $\varepsilon > 0$, there exist $x_\varepsilon, p_\varepsilon$ in \mathbf{R}^n and $q_\varepsilon < 0$ such that*

$$\|x_\varepsilon - x_0\| \leq \varepsilon, \quad \|p_\varepsilon - p\| \leq \varepsilon \quad \& \quad (p_\varepsilon, q_\varepsilon) \in [T_{\mathcal{E}_p(\varphi)}(x_\varepsilon, \varphi(x_\varepsilon))]^-.$$

Consider a set-valued map $F : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ with nonempty compact images and define the Hamiltonian H by (12).

LEMMA 4.3. *Consider an extended lower semicontinuous function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$. Assume that F is upper semicontinuous and has nonempty convex compact images on $\text{Dom}(V)$.*

Then the following four statements are equivalent:

- (i) $\forall (t, x) \in \text{Dom}(V)$ with $t < T$ and $\forall (p_t, p_x, q) \in [T_{\mathcal{E}_p(V)}(t, x, V(t, x))]^-$
- (17) $-p_t + H(t, x, -p_x) \geq 0$;
- (ii) For all $(t, x) \in \text{Dom}(V)$ with $t < T$ and for all $y \geq V(t, x)$
- $(\{1\} \times F(t, x) \times \{0\}) \cap T_{\mathcal{E}_p(V)}(t, x, y) \neq \emptyset$;
- (iii) For all $(t, x) \in \text{Dom}(V)$ with $t < T$

$$\min_{v \in F(t, x)} D_{\uparrow} V(t, x)(1, v) \leq 0;$$

- (iv) For all $(t, x) \in \text{Dom}(V)$ with $t < T$

$$\forall (p_t, p_x) \in \partial_- V(t, x), \quad -p_t + H(t, x, -p_x) \geq 0.$$

Proof. By (14), (ii) \iff (iii). Clearly (ii) \implies (i).

If (i) holds true, then, from the separation theorem,

$$(18) \quad (\{1\} \times F(t, x) \times \{0\}) \cap \overline{\text{co}}(T_{\mathcal{E}_p(V)}(t, x, y)) \neq \emptyset$$

for all $(t, x) \in \text{Dom}(V)$ with $t < T$ and $y \geq V(t, x)$. Finally, since F is upper semicontinuous and has convex compact images, by [21] (see also [6] for a better proof), (18) implies (ii).

By Proposition 4.1, (i) yields (iv). Assume next that (iv) is verified. Fix $(t, x) \in \text{Dom}(V)$ with $t < T$ and $(p_t, p_x, q) \in [T_{\mathcal{E}_p(V)}(t, x, V(t, x))]^-$. Then $q \leq 0$. If $q < 0$, then

$$\left(\frac{p_t}{|q|}, \frac{p_x}{|q|}, -1 \right) \in [T_{\mathcal{E}_p(V)}(t, x, V(t, x))]^-.$$

From Proposition 4.1 and (iv) we deduce (17). It remains to consider the case $q = 0$ and $(p_t, p_x) \neq 0$. For this purpose it is enough to apply Lemma 4.2 and to use the upper semicontinuity of H . \square

LEMMA 4.4. *Consider an extended lower semicontinuous function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ and assume that F is lower semicontinuous and has nonempty compact images on $\text{Dom}(V)$.*

Then the following four statements are equivalent:

(i) $\forall (t, x) \in \text{Dom}(V)$ with $t > 0$ and $\forall (p_t, p_x, q) \in [T_{\mathcal{E}p(V)}(t, x, V(t, x))]^-$

$$-p_t + H(t, x, -p_x) \leq 0;$$

(ii) For all $(t, x) \in \text{Dom}(V)$ with $t > 0$ and for all $y \geq V(t, x)$

$$\{-1\} \times (-F(t, x)) \times \{0\} \subset T_{\mathcal{E}p(V)}(t, x, y);$$

(iii) For all $(t, x) \in \text{Dom}(V)$ with $t > 0$

$$\sup_{v \in F(t, x)} D_{\uparrow} V(t, x)(-1, -v) \leq 0;$$

(iv) For all $(t, x) \in \text{Dom}(V)$ with $t > 0$

$$\forall (p_t, p_x) \in \partial_- V(t, x), \quad -p_t + H(t, x, -p_x) \leq 0.$$

Proof. We deduce from (14) that (ii) is equivalent to (iii). Clearly, (ii) \implies (i). The separation theorem and (i) yield

$$(19) \quad \{-1\} \times (-F(t, x)) \times \{0\} \subset \overline{\text{co}}(T_{\mathcal{E}p(V)}(t, x, y))$$

for all $(t, x) \in \text{Dom}(V)$ with $t > 0$ and $y \geq V(t, x)$.

Since F is lower semicontinuous, (19) and [5, p. 130] imply that for all $(t, x) \in \text{Dom}(V)$ with $t > 0$ and all $y \geq V(t, x)$,

$$\begin{aligned} \{-1\} \times (-F(t, x)) \times \{0\} &\subset \text{Liminf}_{(t', x', y') \rightarrow_{\mathcal{E}p(V)}(t, x, y)} \overline{\text{co}}(T_{\mathcal{E}p(V)}(t', x', y')) \\ &\subset T_{\mathcal{E}p(V)}(t, x, y), \end{aligned}$$

where Liminf denotes the *lower set limit* and $\rightarrow_{\mathcal{E}p(V)}$ the convergence in the epigraph of V . Hence (ii) follows from (i). Arguments similar to those of the proof of Lemma 4.3 yield (i) \iff (iv). \square

5. Alternative boundary conditions. We observe that in Theorem 2.3 (iv) \implies (v), thanks to the equivalence (iv) \iff (i) and in view of Proposition 2.1. In this section we prove that (v) yields (iii). Consider a set-valued map $F : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$.

THEOREM 5.1. *Assume (9). If an extended lower semicontinuous function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ satisfies*

$$\begin{aligned} \forall (t, x) \in]0, T[\times \mathbf{R}^n, \quad \forall (p_t, p_x) \in \partial_- V(t, x), \quad -p_t + H(t, x, -p_x) &= 0, \\ \forall \bar{x} \in \mathbf{R}^n, \quad V(0, \bar{x}) &= \liminf_{t \rightarrow 0+, x \rightarrow \bar{x}} V(t, x), \\ \forall \bar{x} \in \mathbf{R}^n, \quad V(T, \bar{x}) &= \liminf_{t \rightarrow T-, x \rightarrow \bar{x}} V(t, x) \end{aligned}$$

then for all $(t, x) \in \text{Dom}(V)$,

$$(20) \quad \begin{aligned} 0 < t \leq T &\implies \sup_{v \in F(t, x)} D_{\uparrow} V(t, x)(-1, -v) \leq 0, \\ 0 \leq t < T &\implies \inf_{v \in F(t, x)} D_{\uparrow} V(t, x)(1, v) \leq 0. \end{aligned}$$

Proof. From the proofs of Lemmas 4.3 and 4.4 we deduce that for all $(t, x) \in \text{Dom}(V)$ with $0 < t < T$ we have

$$\inf_{v \in F(t, x)} D_{\uparrow} V(t, x)(1, v) \leq 0, \quad \sup_{v \in F(t, x)} D_{\uparrow} V(t, x)(-1, -v) \leq 0.$$

This and Theorems 3.2 and 3.3 imply that for all $0 < t_1 \leq t_2 < T$

$$(21) \quad \forall x \in S_{[t_1, t_2]}(x_1), \quad V(t_1, x_1) \leq V(t_2, x(t_2))$$

and

$$(22) \quad \forall (t_1, x_1) \in \text{Dom}(V), \quad \exists x \in S_{[t_1, t_2]}(x_1) \quad \text{with} \quad V(t_1, x_1) = V(t_2, x(t_2)).$$

We show next that (21) holds true also for $t_2 = T$. Consider any $(t_0, x_0) \in \text{Dom}(V)$ and $x \in S_{[t_0, T]}(x_0)$. Let $x(T) = \bar{x}$ and $t_i \rightarrow T-$, $x_i \rightarrow \bar{x}$ be such that

$$\lim_{i \rightarrow \infty} V(t_i, x_i) = V(T, \bar{x}).$$

By Filippov's theorem [4, p. 120], there exist \bar{y}_i and $y_i \in S_{[0, T]}(\bar{y}_i)$ such that $y_i(t_i) = x_i$ and y_i converge to x uniformly on $[t_0, T]$. Then for all arbitrary but fixed $t_0 < t < T$ and all i large enough, $V(t, y_i(t)) \leq V(t_i, x_i)$. Since V is lower semicontinuous,

$$V(t, x(t)) \leq \liminf_{i \rightarrow \infty} V(t, y_i(t)) \leq \lim_{i \rightarrow \infty} V(t_i, x_i) = V(T, \bar{x}).$$

This, (21), and Theorem 3.3 yield the first inequality in (20).

To prove that (22) holds true also with $t_1 = 0$, fix $(0, \bar{x}) \in \text{Dom}(V)$ and consider $t_i \rightarrow 0+$, $\bar{x}_i \rightarrow \bar{x}$ satisfying

$$V(0, \bar{x}) = \lim_{i \rightarrow \infty} V(t_i, \bar{x}_i).$$

Let $\bar{y}_i \in \mathbf{R}^n$ and $x_i \in S_{[0, T]}(\bar{y}_i)$ be such that $x_i(t_i) = \bar{x}_i$ and

$$\forall t_i \leq t \leq T - \frac{1}{i}, \quad V(t_i, \bar{x}_i) = V(t, x_i(t)).$$

Taking a subsequence and keeping the same notation, by the convergence theorem [5, p. 273], we may assume that x_i converge uniformly to some $x \in S_{[0, T]}(\bar{x})$. Then for all $0 < t < T$,

$$V(0, \bar{x}) = \lim_{i \rightarrow \infty} V(t, x_i(t)) \geq V(t, x(t)).$$

This, (22), and Theorem 3.2 imply the second inequality in (20). \square

6. Comparisons with viscosity solutions. Let $F : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ be a set-valued map with nonempty compact images and H be given by (12). Consider the Hamilton–Jacobi–Bellman equation

$$(23) \quad -\frac{\partial V}{\partial t}(t, x) + H\left(t, x, -\frac{\partial V}{\partial x}(t, x)\right) = 0$$

An extended lower semicontinuous function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ is called a *viscosity supersolution* of (23) if for all $(t, x) \in \text{Dom}(V)$ with $t \in]0, T[$ we have

$$\forall (p_t, p_x) \in \partial_- V(t, x), \quad -p_t + H(t, x, -p_x) \geq 0.$$

Remark. We refer to [12], [14] for the continuous case and also underline that the interior of $\text{Dom}(V)$ may be empty. This leads to a slight difference with the usual definition of viscosity supersolutions, usually defined on the closure of an open set.

However, for optimal control problems with constraints it may happen that the domain of definition of the value function has an empty interior. For this reason it is more natural to deal with arbitrary lower semicontinuous solutions of HJB equations.

In particular, any V satisfying (v) of Theorem 2.3 is a viscosity supersolution.

THEOREM 6.1. *Let $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semicontinuous function. Assume that F is upper semicontinuous and has nonempty convex compact images on $\text{Dom}(V)$.*

Then the following two statements are equivalent:

- (i) V is a viscosity supersolution of (23);
- (ii) For all $0 < t < T$ and $x \in \mathbf{R}^n$ such that $V(t, x) \neq +\infty$, we have

$$(24) \quad \inf_{v \in F(t, x)} D_{\uparrow} V(t, x)(1, v) \leq 0.$$

The proof follows by the same arguments as in the proof of Lemma 4.3.

Next we recall the notion of viscosity subsolution of (23).

DEFINITION 6.2. Consider an extended function $\varphi : \mathbf{R}^n \mapsto \mathbf{R} \cup \{\pm\infty\}$. The superdifferential of φ at $x_0 \in \text{Dom}(\varphi)$ is defined by

$$p \in \partial_+ \varphi(x_0) \iff \limsup_{x \rightarrow x_0} \frac{\varphi(x) - \varphi(x_0) - \langle p, x - x_0 \rangle}{\|x - x_0\|} \leq 0.$$

The contingent *hypoderivative* of φ at $x_0 \in \text{Dom}(\varphi)$ in the direction $u \in \mathbf{R}^n$ is defined by

$$D_{\downarrow} \varphi(x_0)(u) = \limsup_{h \rightarrow 0+, u' \rightarrow u} \frac{\varphi(x_0 + hu') - \varphi(x_0)}{h}.$$

Clearly,

$$\partial_+ \varphi(x_0) = -\partial_-(-\varphi)(x_0) \quad \text{and} \quad D_{\downarrow} \varphi(x_0)(u) = -D_{\uparrow}(-\varphi)(x_0)(u)$$

and

$$T_{\mathcal{Hyp}(\varphi)}(x_0, \varphi(x_0)) = \mathcal{Hyp}(D_{\downarrow} \varphi(x_0)),$$

where \mathcal{Hyp} states for the hypograph. In particular, $p \in \partial_+ \varphi(x_0)$ if and only if

$$(25) \quad \forall u \in \mathbf{R}^n, \quad D_{\downarrow} \varphi(x_0)(u) \leq \langle p, u \rangle.$$

Thus

$$(26) \quad p \in \partial_+ \varphi(x_0) \iff (-p, +1) \in (T_{\mathcal{Hyp}(\varphi)}(x_0, \varphi(x_0)))^-.$$

An extended upper semicontinuous function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{-\infty\}$ is called a *viscosity subsolution* of (23) if for all $(t, x) \in \text{Dom}(V)$ with $0 < t < T$ we have

$$\forall (p_t, p_x) \in \partial_+ V(t, x), \quad -p_t + H(t, x, -p_x) \leq 0.$$

In the above we set $V = -\infty$ on the complement of $[0, T] \times \mathbf{R}^n$ and define the superdifferential for such extended function.

THEOREM 6.3. *Let $V : [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R} \cup \{\pm\infty\}$ be an extended function whose domain is closed and such that the restriction of V to $\text{Dom}(V)$ is continuous. Assume that F satisfies (9). Then the following two statements are equivalent:*

- (i) V is a viscosity subsolution of (23);
- (ii) For all $0 < t < T$ and all $x \in \mathbf{R}^n$ with $(t, x) \in \text{Dom}(V)$,

$$\sup_{v \in F(t, x)} D_{\uparrow} V(t, x)(-1, -v) \leq 0.$$

Proof. Assume (ii) and set $V(t, x) = +\infty$ for all $(t, x) \notin \text{Dom}(V)$. Then V is lower semicontinuous. Fix $0 < t_0 < T$. By Theorem 3.3, for every $t_0 \leq t_1 < T$ and every $x \in S_{[t_0, t_1]}(x_0)$ the following holds true:

$$\forall t \in [t_0, t_1], \quad V(t_0, x_0) \leq V(t, x(t))$$

Fix $v \in F(t_0, x_0)$. From Filippov's theorem [4, p. 120], there exist $t_0 < t_1 < T$ and $x \in S_{[t_0, t_1]}(x_0)$ such that $x'(t_0) = v$. The above inequality yields $0 \leq D_{\downarrow} V(t_0, x_0)(1, v)$. Consequently,

$$\forall (p_t, p_x) \in \partial_+ V(t_0, x_0), \quad 0 \leq p_t + \langle p_x, v \rangle$$

Since $v \in F(t_0, x_0)$ is arbitrary, V is a viscosity subsolution.

Assume next that (i) is verified. We set $V(t, x) = -\infty$ for all $(t, x) \notin \text{Dom}(V)$. Then the hypograph $\mathcal{Hyp}(V)$ of V is closed. Exactly as at the end of the proof of Lemma 4.3 we show that for all $(t, x) \in \text{Dom}(V)$ with $0 < t < T$ and all $z \leq V(t, x)$

$$(27) \quad \forall (q_t, q_x, q) \in (T_{\mathcal{Hyp}(V)}(t, x, z))^{-}, \quad q_t + H(t, x, q_x) \leq 0.$$

We next deduce from (27) and the separation theorem that for all $(t, x) \in \text{Dom}(V)$ with $0 < t < T$ and all $z \leq V(t, x)$

$$\{1\} \times F(t, x) \times \{0\} \subset \overline{co}(T_{\mathcal{Hyp}(V)}(t, x, z)).$$

This, [5, p. 130], and lower semicontinuity of F imply that for all $(t, x) \in \text{Dom}(V)$ with $0 < t < T$

$$\begin{aligned} \{1\} \times F(t, x) \times \{0\} &\subset \liminf_{\substack{(t', x', z') \rightarrow (t, x, V(t, x)) \\ (t', x', z') \in \mathcal{Hyp}(V)}} \overline{co}(T_{\mathcal{Hyp}(V)}(t', x', z')), \\ &\subset T_{\mathcal{Hyp}(V)}(t, x, V(t, x)) = \mathcal{Hyp}(D_{\downarrow} V(t, x)). \end{aligned}$$

Thus for all $(t, x) \in \text{Dom}(V)$ satisfying $0 < t < T$,

$$\inf_{v \in F(t, x)} D_{\downarrow} V(t, x)(1, v) \geq 0.$$

Consider the extended lower semicontinuous $W : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ defined by $W(t, x) = -V(T - t, x)$. Then for all $(t, x) \in \text{Dom}(W)$ with $0 < t < T$ and all $v \in F(T - t, x)$, we have

$$D_{\uparrow} W(t, x)(-1, v) = -D_{\downarrow} V(T - t, x)(1, v) \leq 0.$$

Fix any $(t_0, x_0) \in \text{Dom}(V)$ with $0 < t_0 < T$, $v \in F(t_0, x_0)$ and consider a solution $y(\cdot)$ of the differential inclusion

$$\begin{aligned} y' &\in -F(T - t, y), \\ y(T - t_0) &= x_0, \quad y'(T - t_0) = -v. \end{aligned}$$

(It exists by [4, p. 120].) Then, applying Theorem 3.3 with W and the set-valued map $(t, x) \rightsquigarrow -F(T - t, x)$, we deduce that for all small $s > 0$,

$$W(T - t_0, x_0) \leq W(T - t_0 + s, y(T - t_0 + s))$$

and therefore for a sequence $v_s \rightarrow v$ we have $V(t_0 - s, x_0 - sv_s) \leq V(t_0, x_0)$. Thus $D_1 V(t_0, x_0)(-1, -v) \leq 0$. Since $v \in F(t_0, x_0)$ is arbitrary, (ii) follows. \square

Let $\text{Dom}(V) \subset [0, T] \times \mathbf{R}^n$ be a closed set such that (13) holds true and $V : \text{Dom}(V) \mapsto \mathbf{R}$ be a continuous *viscosity solution* of (23) (i.e., it is simultaneously super and subsolution in the viscosity sense). Then, using Theorems 6.1 and 6.3 and Proposition 4.1, we check that V verifies (v) of Theorem 2.3. Conversely, if V verifies (v), then (iii) holds true and from Theorems 6.1 and 6.3, V is a viscosity solution of (23).

The proof of Theorem 2.3 is completed.

7. Representation of solutions of Hamilton–Jacobi equations with convex Hamiltonians. Consider $H : [0, T] \times \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}$ and the Hamilton–Jacobi–Bellman equation

$$(28) \quad -\frac{\partial V}{\partial t}(t, x) + H\left(t, x, -\frac{\partial V}{\partial x}(t, x)\right) = 0.$$

We look for its solutions in the class of lower semicontinuous extended functions $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$.

In this section we impose the following assumptions:

- (i) H is continuous;
- (ii) $H(t, x, \cdot)$ is convex;
- (29) (iii) $\exists k \in L^1(0, T), \forall p \in B, \|H(t, x, p)\| \leq k(t)(1 + \|x\|)$;
- (iv) $\forall R > 0, \exists c_R \in L^1(0, T)$ such that for almost all $t \in [0, T]$
 $\forall p \in B, H(t, \cdot, p)$ is $c_R(t)$ –Lipschitz on $B_R(0)$;
- (v) $H(t, x, \cdot)$ is positively homogeneous,

where B denotes the closed unit ball in \mathbf{R}^n .

Remark. In all the results of this section we assume (29). However, assumption (v) may be replaced by the Lipschitz continuity of $H(t, x, \cdot)$ together with (modified with respect to p) conditions (iii), (iv). Then it is possible to study solutions of (28) via a Hamilton–Jacobi–Bellman equation with the new (conjugate) Hamiltonian meeting assumptions (29) (as was done, for instance, in [8]). We avoid a detailed discussion on such a more general case, because the optimal control problems give rise to positively homogeneous $H(t, x, \cdot)$.

Define $F : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ by

$$(30) \quad F(t, x) = \bigcap_{\|p\|=1} \{v \in \mathbf{R}^n \mid \langle p, v \rangle \leq H(t, x, p)\}$$

PROPOSITION 7.1. *If (29) holds true, then F verifies (9) and*

$$\forall p \in \mathbf{R}^n, \quad \sup_{v \in F(t, x)} \langle p, v \rangle = H(t, x, p).$$

Proof. By [11, p. 54] F has convex compact images and is continuous. Let $\bar{v} \in F(t, x)$ be such that $\|\bar{v}\| = \sup_{v \in F(t, x)} \|v\|$. Applying assumption (29) (iii) with $p = \bar{v} / \|\bar{v}\|$, we obtain (9) (b). Fix $t \in [0, T]$, $R > 0$ and let $x, y \in B_R(0)$, $z \in F(t, x)$. Observe that for every $p \in B$,

$$\langle p, z \rangle \leq H(t, x, p) \leq H(t, y, p) + c_R(t) \|x - y\|.$$

Hence, by the separation theorem, $z \in F(t, y) + c_R(t) \|x - y\| B$. Thus $F(t, x) \subset F(t, y) + c_R(t) \|x - y\| B$ and (9) (c) follows. The last statement is a well-known result of functional analysis. \square

THEOREM 7.2. *Assume (29) and consider an extended lower semicontinuous function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$.*

Then the following two statements are equivalent:

(i) *V satisfies the equation*

$$(31) \quad \forall (t, x) \in]0, T[\times \mathbf{R}^n, \quad \forall (p_t, p_x) \in \partial_- V(t, x), \quad -p_t + H(t, x, -p_x) = 0$$

and inequalities

$$\forall (p_t, p_x) \in \partial_- V(0, x), \quad -p_t + H(0, x, -p_x) \geq 0$$

$$\forall (p_t, p_x) \in \partial_- V(T, x), \quad -p_t + H(T, x, -p_x) \leq 0;$$

(ii) *V satisfies (31) and*

$$\forall \bar{x} \in \mathbf{R}^n, \quad V(0, \bar{x}) = \liminf_{(t, x) \rightarrow (0+, x)} V(t, x) \quad \text{and} \quad V(T, \bar{x}) = \liminf_{(t, x) \rightarrow (T-, x)} V(t, x).$$

Furthermore, define the extended lower semicontinuous function $g(\cdot) = V(T, \cdot)$ and let F be given by (30). Then the above statements are equivalent to

(iii) *For all $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$,*

$$(32) \quad V(t_0, x_0) = \inf \{g(x(T)) \mid x \in S_{[t_0, T]}(x_0)\}.$$

Finally, if $\text{Dom}(V)$ is closed and the restriction of V to its domain is continuous and if

$$(\{0, T\} \times \mathbf{R}^n) \cap \text{Dom}(V) \subset \overline{[0, T[\times \mathbf{R}^n \cap \text{Dom}(V)},$$

then the above statements are equivalent to the following:

V is a viscosity solution of Hamilton–Jacobi equation (28) on $\text{Dom}(V)$.

The proof follows from Theorem 2.3 and Proposition 7.1.

COROLLARY 7.3 (maximum principle). *Assume (29) and let V_1, V_2 be extended lower semicontinuous functions from $[0, T] \times \mathbf{R}^n$ into $\mathbf{R} \cup \{+\infty\}$ satisfying (i) (or, equivalently, (ii)) of Theorem 7.2.*

If $V_1(T, \cdot) \geq V_2(T, \cdot)$, then $V_1 \geq V_2$.

Proof. By Theorem 7.2, V_i is given by (32) with $g = V_i(T, \cdot)$. \square

Remark. By the analogy with viscosity solutions, an extended lower semicontinuous function $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ can be called a solution of (28) if it verifies (i) or (ii) of Theorem 7.2. In particular, in [8] (ii) is taken as a definition of solution.

8. Stability of value functions. In this section we show that the lower epilimit of value functions is still the value function of a Mayer problem. Definitions and properties of epilimits can be found, for instance, in [5, Chap. 7].

THEOREM 8.1. *Consider set-valued maps $F_i : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$, $i \geq 1$ satisfying (9) with $k(\cdot)$, $c_R(\cdot)$ independent of i , and extended lower semicontinuous functions $g_i : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$. Assume that $\inf_{i \geq 1, x \in \mathbf{R}^n} g_i(x) > -\infty$.*

Let \mathcal{V}_i be defined by (32) with F and g replaced by F_i and g_i , respectively, and \mathcal{V} denote the lower epilimit of \mathcal{V}_i . Set $g(\cdot) = \mathcal{V}(T, \cdot)$.

If for a set-valued map $F : [0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ and for every $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$

$$\lim_{i \rightarrow \infty, (t, x) \rightarrow (t_0, x_0)} F_i(t, x) = F(t_0, x_0),$$

then \mathcal{V} verifies (32).

Proof. We first claim that \mathcal{V} is nondecreasing along solutions of differential inclusion (6). Indeed fix $t_1 \in]t_0, T]$, $x \in S_{[t_0, T]}(x_0)$. Then for almost all $t \in [t_0, T]$,

$$\forall i \geq 1, \quad \text{dist} \left(x'(t), F_i(t, x(t)) \right) \leq \|x'(t)\| + k(t)(1 + \|x(t)\|)$$

$$\lim_{i \rightarrow \infty} \text{dist} \left(x'(t), F_i(t, x(t)) \right) = 0.$$

Let $(t_j, x_1^j) \rightarrow (t_1, x(t_1))$ be such that $\mathcal{V}(t_1, x(t_1)) = \lim_{j \rightarrow \infty} \mathcal{V}_{i_j}(t_j, x_1^j)$. From Filippov's theorem [4, p. 120] there exist solutions x_j of

$$\begin{aligned} x_j'(t) &\in F_{i_j}(t, x_j(t)), \quad t \in [t_0, T], \\ x_j(t_j) &= x_1^j \end{aligned}$$

such that $\lim_{j \rightarrow \infty} \sup_{t \in [t_0, T]} \|x_j(t) - x(t)\| = 0$. Thus, for all large j ,

$$\mathcal{V}_{i_j}(t_0, x_j(t_0)) \leq \mathcal{V}_{i_j}(t_j, x_j(t_j)) = \mathcal{V}_{i_j}(t_j, x_1^j)$$

and from the lower semicontinuity of \mathcal{V} we obtain

$$\mathcal{V}(t_0, x_0) \leq \lim_{j \rightarrow \infty} \mathcal{V}_{i_j}(t_j, x_1^j) = \mathcal{V}(t_1, x(t_1)).$$

We next show that for all $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$, there exists $x \in S_{[t_0, T]}(x_0)$ such that $\mathcal{V}(t_0, x_0) \geq g(x(T))$. Indeed fix $(t_0, x_0) \in \text{Dom}(\mathcal{V})$ and let $(t_j, x_j) \rightarrow (t_0, x_0)$ be such that

$$\lim_{j \rightarrow \infty} \mathcal{V}_{i_j}(t_j, x_j) = \mathcal{V}(t_0, x_0).$$

Consider $\bar{x}_j \in S_{[t_j, T]}(x_j)$ such that $\mathcal{V}_{i_j}(t_j, x_j) = g_{i_j}(\bar{x}_j(T))$. Taking a subsequence and keeping the same notations we may assume that \bar{x}_j converge (pointwise) to some $\bar{x} \in S_{[t_0, T]}(x_0)$. Thus, by the lower semicontinuity of g ,

$$\mathcal{V}(t_0, x_0) = \lim_{j \rightarrow \infty} g_{i_j}(\bar{x}_j(T)) \geq g(\bar{x}(T)) = V(T, \bar{x}(T))$$

From the equivalence of i) and ii) in Theorem 2.3 we deduce that \mathcal{V} is the value function of problem (7). \square

Acknowledgments. I would like to thank P.-L. Lions, who brought my attention to [8], [9] while I was preparing this manuscript and the unknown referee for many constructive suggestions, which helped to improve the presentation of results.

REFERENCES

- [1] J.-P. AUBIN, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusions*, Adv. in Math. Suppl. Stud., L. Nachbin, ed., Academic Press, New York, 1981, pp. 160–232.
- [2] ———, *A survey of viability theory*, SIAM J. Control Optim., 28(1990), pp. 749–788.
- [3] ———, *Viability Theory*, Birkhäuser, Boston, Basel, Berlin, 1991.
- [4] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, New York, Berlin, 1984.
- [5] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, Basel, Berlin, 1990.
- [6] ———, *Partial differential inclusions governing feedback controls*, IIASA WP-90-28, 1990.
- [7] ———, *Inclusions aux dérivés partielles gouvernant des contrôles de rétroaction*, C. R. Acad. Sci. Paris Sér. I Math, 311 (1990), pp. 851–856.
- [8] E. N. BARRON AND R. JENSEN, *Semicontinuous Viscosity Solutions for Hamilton–Jacobi Equations with Convex Hamiltonians*, Preprint, March 1990.
- [9] ———, *Optimal Control and Semicontinuous Viscosity Solutions*, Preprint, March 1990.
- [10] P. CANNARSA AND H. FRANKOWSKA, *Some characterizations of optimal trajectories in control theory*, SIAM J. Control Optim., 29(1991), pp. 1322–1347.
- [11] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math., Vol. 580, Springer-Verlag, New York, Berlin, 1977.
- [12] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277(1983), pp. 1–42.
- [13] ———, *On the existence and uniqueness of solutions of Hamilton–Jacobi equations*, J. Nonlinear Anal., 10(1989), pp. 353–370.
- [14] M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282(1984), pp. 487–502.
- [15] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, Heidelberg, New York, 1975.
- [16] H. FRANKOWSKA, *L'équation d'Hamilton–Jacobi contingente*, C. R. Acad. Sci. Paris Sér. I Math., 304(1987), pp. 295–298.
- [17] ———, *Optimal trajectories associated to a solution of contingent Hamilton–Jacobi equations*, in Proc. IEEE, 26th CDC Conference, Los Angeles, December 9–11, 1987.
- [18] ———, *Non smooth solutions to an Hamilton–Jacobi equation*, in Modeling and Control of Systems in Engineering, Quantum Mechanics, Economics, and Biosciences, A. Blaquiere, ed., Lecture Notes in Control and Inform. Sci., Vol. 121, Springer-Verlag, New York, Berlin, 1988.
- [19] ———, *Optimal trajectories associated to a solution of contingent Hamilton–Jacobi equations*, Appl. Math. Optim., 19(1989), pp. 291–311.
- [20] ———, *Hamilton–Jacobi equation: Viscosity solutions and generalized gradients*, J. Math. Anal. Appl., 141(1989), pp. 21–26.
- [21] H. G. GUSEINOV, A. I. SUBBOTIN, AND V. N. USHAKOV, *Derivatives for multivalued mappings with applications to game-theoretical problems of control*, Problems Control and Inform., 14 (1985), pp. 155–168.
- [22] G. HADDAD, *Monotone trajectories of differential inclusions with memory*, Israel J. Math., 39 (1981), pp. 83–100.
- [23] A. D. IOFFE, *Proximal analysis and approximate subdifferentials*, J. London Math. Soc., 41(1990), pp. 175–192.
- [24] R. T. ROCKAFELLAR, *Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res., 6(1981), pp. 424–436.
- [25] A. I. SUBBOTIN, *A generalization of the basic equation of the theory of differentials games*, Soviet. Math. Dokl., 22(1980), pp. 358–362.
- [26] ———, *Generalization of the main equation of differentials game theory*, J. Optim. Theory Appl., 43(1984), pp. 103–133.

A TRIBUTE TO WENDELL H. FLEMING

For forty years Wendell Fleming has advanced mathematics, sometimes in fundamental ways, in the areas of geometric measure theory, the calculus of variations, the theory of differential games, stochastic control theory, and population genetics. In addition, he has served as an educator, an administrator, and a national leader in policy matters affecting mathematics. His dedication to mathematics and his tireless support of mathematicians, especially young ones, has been a hallmark throughout his long, illustrious career. Those of us who have had the good fortune to know and work with Wendell have come to admire his wisdom, integrity, and kindness, and to cherish his friendship.

Wendell Fleming was born March 7, 1928, in Guthrie, Oklahoma, where his father was teaching temporarily. In 1929 the family returned to Indiana, where Fleming, who considers himself a native Hoosier, grew up on a farm near Sullivan. Upon graduation from Sullivan High School in 1945, he entered Purdue University, majoring in chemical engineering. There he met Flo Tatum, whom he married in 1948. At that time Purdue had (and still has) a course designed to give students insight into their future careers in engineering. After taking this course, Fleming decided to switch to mathematics. He attributes this decision to the discovery that he was “good at math.” One of his chemical engineering professors, upon learning of his decision, told him that mathematics was a stable subject—nothing had changed in 100 years and nothing would change in the next 100 years. After graduating from Purdue, Fleming attended the University of Wisconsin, where he wrote a thesis in surface area theory, under the direction of L. C. Young. He received his Ph.D. from Wisconsin in 1951.

For the next four years, Fleming was employed as a mathematician by the Rand Corporation. Returning to academic life in 1955, Fleming accepted a position as assistant professor at Purdue University. He moved to Brown University in 1958, and has remained on the Brown faculty. He served a term as chairman of the Department of Mathematics and is now in his third term as chairman of the Division of Applied Mathematics. Since joining Brown, he has spent a year at the University of Wisconsin, a year at Stanford University, a semester at the Massachusetts Institute of Technology, a semester at the Institute for Mathematics and its Applications at the University of Minnesota, and shorter periods of time at other institutions.

Fleming was granted a National Science Foundation Fellowship for the period 1968–1969 and was named a Guggenheim Fellow for 1976–1977. He was an invited plenary speaker at the 1983 International Congress of Mathematicians. In 1987 he was awarded the American Mathematical Society’s Steele Prize for his work with Herbert Federer on geometric measure theory. In 1991 Fleming was awarded the degree Doctor of Science, *Honoris Causa*, by Purdue University.

Wendell Fleming has authored or coauthored three books. His first, *Functions of Several Variables*, is a highly regarded undergraduate text. *Deterministic and Stochastic Optimal Control*, coauthored with Raymond Rishel, has been a standard reference in stochastic optimal control for fifteen years. The recent development of viscosity solutions of partial differential equations has inspired a new book on stochastic optimal control, entitled *Controlled Markov Processes and Viscosity Solutions*, written with Mete Soner.

Twenty students, including the third author of this tribute, have obtained Ph.D. degrees under Wendell Fleming’s supervision. Many of these students are active in

research, and work of two of them, Mete Soner and Thaleia Zariphopoulou, appears in this issue.

Good citizenship has been a high priority in Wendell Fleming's career. He has served on the editorial boards of eight journals, including the *SIAM Journal on Control and Optimization* from 1979–1990, and as an editor for two book series. He has served on numerous committees of the Conference Board of Mathematical Sciences and the American Mathematical Society, and for five years compiled reports on graduate education and employment for the *AMS Notices*. During 1981–1984 he was a member of the Board of Governors of the Institute for Mathematics and its Applications (IMA), and he helped organize the IMA programs for the years 1985–1986 and 1992–1993. From 1986–1988 he chaired the Air Force Review Panel for Mathematical Sciences and the Panel on Future Directions in Control Theory, from which emerged an influential report.

Wendell Fleming's research to date can be roughly divided into four areas: (a) the calculus of variations and geometric measure theory, (b) differential games, (c) stochastic control, and (d) population genetics. Each of these is discussed below.

(a) Beginning in the 1950s, Fleming played a major role in both the rapid development of the calculus of variations and the initiation of the new field of geometric measure theory. He made many significant and lasting contributions to these subjects, including some that were fundamental. L. C. Young, Fleming's Ph.D. supervisor, introduced the theory of generalized surfaces for the purpose of creating a general setting for problems in multi-dimensional calculus of variations. Both Fleming and Young made great strides in developing the theory of generalized surfaces and this work undoubtedly influenced Fleming's subsequent work with Federer.

In 1960, Fleming coauthored a seminal paper with Herbert Federer, "Normal and integral currents," *Ann. of Math.*, 72 (1960), pp. 458–520, which marked the beginning of geometric measure theory as it is known today. This paper introduced and developed the theory of integral currents, which, like generalized surfaces, provides a general framework for geometric problems in the calculus of variations. De Rham's concept of a k -dimensional current in R^n is used to identify a k -dimensional oriented manifold M with the operation of integrating a k -form over M . De Rham's currents are linear functionals on differential k -forms and become Schwartz distributions in the case $k = 0$. To be useful in applications, integral currents had to be general enough to allow integration of forms over sets that differed in small measure from oriented C^1 -manifolds, while at the same time they had to provide compactness properties needed for existence of solutions to geometric problems in the calculus of variations, such as the problem of least area (Plateau's Problem) in arbitrary dimension and codimension. In the 1930s, Douglas and Rado were the first to make substantial progress toward the resolution of Plateau's problem. Their work established the existence of a solution to the problem of finding a surface of least area of prescribed topological type bounded by a given curve in R^3 . However, their methods were not general enough to treat the phenomenon illustrated by Fleming in his paper "An example in the problem of least area," *Proc. Amer. Math. Soc.*, 7 (1956), pp. 1063–1074. Here he showed the existence of a simple closed rectifiable curve in R^3 such that the problem of least area with unrestricted topological type has no solution of finite topological type. The theory of integral currents is general enough to accommodate this pathology. It also supplies a powerful linkage between measure theory and algebraic topology and gives the existence of minimizing integral currents in integral homology classes. For their remarkable achievements, Federer and Fleming were awarded the Steele Prize, which carried with it the following citation: "The 1987 Steele Prize for a paper which has

proved to be of fundamental or lasting importance in its field is awarded to Herbert Federer and Wendell Fleming for their pioneering paper "Normal and integral currents".

Fleming used the theory of integral currents to investigate the regularity of solutions to the problem of least area in the paper "On the oriented Plateau problem," *Rend. Circ. Mat. Palermo*, 11 (1962), pp. 1-22. Previous attempts on the problem of least area, including those by Douglas and Rado mentioned earlier, did not completely resolve the question of the solution's regularity, allowing the possibility of branch points. Fleming settled the question in the case of a two-dimensional oriented minimizing current in R^3 by showing that, away from the boundary, it is a smooth minimal surface in the sense of classical differential geometry. Techniques developed in this paper have played a key role in subsequent work on regularity of minimal currents. In this paper Fleming also provided a completely new method of attacking the Bernstein conjecture, which states that a smooth function satisfying the minimal surface equation on R^{n-1} has the property that its graph is a hyperplane in R^n . At that time, proofs were known for the case $n = 3$, and most used methods of complex analysis. Fleming's proof did not. He showed that the Bernstein conjecture would follow from an interior regularity result for minimal integral currents with codimension one in R^n . His regularity theorem thus gave another proof of the Bernstein conjecture for $n = 3$. His discovery concerning regularity results for minimal currents and the Bernstein conjecture was used extensively in subsequent efforts by various authors that eventually led to a complete resolution of the conjecture.

(b) In the early 1960s Fleming published two papers that greatly influenced developments in differential game theory. The study of differential games was initiated in the mid 1950s by Isaacs, who treated many examples in a formal fashion. A mathematically satisfactory general treatment of the notion of strategy and of the question of the existence of value had, however, not been developed. The first major step in this direction was Fleming's 1961 paper, "The convergence problem for differential games," *J. Math. Anal. Appl.*, 3 (1961), pp. 102-116. Fleming discretized time and replaced the differential equation dynamics by difference equation dynamics. The integral payoff was also replaced by a discrete approximation. Decisions were made by each player at each of n discrete times, with full knowledge of the history of the game prior to, but not including, the current time. Fleming showed that the values $V_n(x, T)$ of the n th step discrete game tended to a limit $V(x, T)$ as $n \rightarrow \infty$. This limit he took to be the value of the game. To prove this result, Fleming introduced the discrete-time majorant (or upper) game and the discrete-time minorant (or lower) game, in which the maximizing and minimizing players, respectively, have information advantages. In this paper certain restrictive assumptions were made concerning the form of the dynamics and the payoff.

In his second paper, "The convergence problem for differential games II," in *Advances in Game Theory*, *Ann. of Math. Stud.*, Vol. 52, M. Dresher, L. S. Shapley, and A. W. Tucker, eds., Princeton University Press, 1964, Fleming removed the restrictions on the dynamics and on the integrand of the payoff. To achieve this, he introduced the novel device of considering the discrete game with a small noise term added at each decision time. He showed that if $\varepsilon > 0$ is the measure of the smallness of the noise and if $V_n^\varepsilon(x, T)$ denotes the value of the n th stage discrete game with noise term, then for each fixed $\varepsilon > 0$, $V_n^\varepsilon(x, T)$ tends to a function $V^\varepsilon(x, T)$ as $n \rightarrow \infty$. Here $V^\varepsilon(x, T)$ is the solution of the Isaacs equation with a term $(\varepsilon^2/2)\Delta V$ added, where Δ denotes the Laplacian. He also showed that if $V_n(x, T)$ is, as before, the value of the n th stage game without noise, then $|V_n^\varepsilon(x, T) - V_n(x, T)|$ is uniformly small. From

this it follows that $V_n(x, T)$ tends to a limit, which he again defined to be the value of the game. For the next 20 years the ideas introduced in these papers were taken up by many authors in their treatments of differential games. The partial differential equation considerations connected with small-noise games were precursors of the development of the theory of viscosity solutions, whose recent introduction has greatly simplified the theory and removed the need to study small-noise games.

(c) In the late 1960s, Wendell Fleming began pioneering work in the formulation and study of continuous-time, nonlinear, stochastic control models. He first provided conditions under which models with complete observations and models with partial observations were well-posed, raised and offered some resolution to the thorny problem of degeneracy of the underlying diffusion, and broached the issue of singular perturbation of deterministic problems by the addition of a small-noise term. This work was centered around the Hamilton–Jacob–Bellman equation, and relied on the author’s facility with nonlinear partial differential equations as well as stochastic analysis. With the publication of “Optimal control of partially observable diffusions,” *SIAM J. Control Optim.*, 6 (1968), pp. 194–214, Fleming initiated the study of continuous-time nonlinear problems with partial observations, providing a modeling framework and methods for the proof of existence of optimal controls. The paper “Duality and a priori estimates in Markovian optimization problems,” *J. Math. Anal. Appl.*, 16 (1966), pp. 254–279, introduces nonanticipative controls as a device to overcome the complications arising from the nonsmooth dependence of the control variable on the state variable. This paper also shows how to characterize the solution of a degenerate Hamilton–Jacobi–Bellman equation as the unique almost everywhere solution obtained as a limit of solutions to perturbed problems. Such equations typically have many almost everywhere solutions; Fleming’s method is now understood to yield the “viscosity solution,” which is the one of interest in control theory. In “Stochastic control for small noise intensities,” *SIAM J. Control Optim.*, 9 (1971), pp. 473–517, Fleming obtained an expansion of the value function in powers of a noise parameter. This kind of analysis, which contains a first glimmer of large deviation theory for diffusion processes, has been simplified and expanded through the use of viscosity solutions; Fleming has been intimately involved in these developments.

(d) In the 1970s, Wendell Fleming turned his considerable analytical skills to problems in population genetics. Learning the models and issues of population genetics, he contributed several highly regarded papers.

A central issue in population biology regards the existence and the character of equilibrium distributions of genetic types (alleles). Fleming’s first paper in the field, coauthored with Chau-Hsing Su, was “Some one-dimensional migration models in population genetics theory,” *Theoret. Population Biol.*, 5 (1974), pp. 431–449. Fleming and Su investigated mutation, random genetic drift (the stochastic fluctuation of gene frequencies in finite populations), and migration in a line segment. They formulated a boundary value problem for the covariance between the gene frequencies at two arbitrary points in the habitat, solved it at equilibrium, and found the eigenvalues and eigenfunctions that control convergence to equilibrium.

In “A selection-migration model in population genetics,” *J. Math. Biol.*, 2 (1975), pp. 219–233, Fleming analyzed a partial differential equation model for gene frequency in a bounded, one-dimensional habitat. Fleming’s model, a generalization of the classical Fisher model and closely tied to contemporaneous research activity, provides a sterling example of careful and complete analysis with genuine applied significance. Using Lyapunov function techniques to study the asymptotic behavior of the solution

to the equation at hand, Fleming identified both stable and unstable equilibria in different parameter regimes and discovered bifurcation of equilibria. Moreover, he showed in this general framework how extinction of one allele could occur, a somewhat surprising result.

In "Equilibrium distributions of continuous polygenic traits," *SIAM J. Appl. Math.*, 36 (1979), pp. 148–168, Fleming undertook to model the distribution of a continuous trait (such as size) influenced by a finite number of loci. The evolutionary processes involved were selection, mutation, and recombination (paternal and maternal genes mixing to form offspring). Again, one wants to know the equilibrium distribution. The existing models at the time either assumed independence of the distributions at different loci or else assumed a multivariate normal equilibrium distribution at the loci under consideration. Instead of making these kinds of restrictive assumptions, Fleming chose to exploit the fact that mutation and selection are weak influences, and to expand the equilibrium distribution in terms of these small parameters. The expansion of the equilibrium he obtained turned out to have the normal distribution as its leading term, but also nonnormal deviations introduced by higher-order terms.

Wendell Fleming's best-known work in population genetics is the fundamental paper "Some measure-valued Markov processes in population genetics theory," *Indiana Univ. Math. J.*, 28 (1979), pp. 817–843, with Michel Viot. Fleming and Viot constructed a measure-valued process to model populations subject to mutation, natural selection, and random genetic drift. Prior to the construction of this so-called *Fleming–Viot process*, population geneticists were restricted to models in which traits or "types" of individuals were discrete, so that the distribution of types in a population could be described as a finite-dimensional, or perhaps countably-infinite-dimensional, probability vector. Using approximation arguments and the martingale characterization of Markov processes, Fleming and Viot managed to build a process taking values in the set of measures on a locally compact space of types. Despite the generality of the Fleming–Viot process, its usefulness, even at the level of explicit computations, has been subsequently demonstrated by the work of many authors. The reader is referred to the survey paper "Fleming–Viot processes in population genetics," by T. Kurtz and S. Ethier, appearing in this issue.

On behalf of the numerous mathematicians and other scientists who have been inspired by the work and person of Wendell Fleming, we express to him our gratitude through the dedication of this special issue of the *SIAM Journal on Control and Optimization* on the occasion of Wendell's sixty-fifth birthday.

L. D. Berkovitz

Purdue University

Steven E. Shreve

Carnegie Mellon University

William P. Ziemer

Indiana University

Editor's Note

The SIAM staff is indebted to everyone who made this issue possible. In particular, we would like to thank Steven Shreve, of Carnegie Mellon University, for serving as Editor-in-Chief for this issue. We would also like to thank the following editors for their efforts: Leonard Berkovitz (Purdue University); Thomas Kurtz (University of Wisconsin); Pierre-Louis Lions (Université du Paris-Dauphine); and Etienne Pardoux

(Université de Provence). Thanks are also due to Wendell Fleming for providing a list of his publications and the photo that accompanies this tribute.

PUBLICATIONS BY WENDELL H. FLEMING

BOOKS

- Functions of Several Variables*, Addison-Wesley, Reading, MA, 1965; 2nd ed., Springer-Verlag, New York, Berlin, 1977.
- with R. W. Rishell, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, Berlin, 1975.
- co-editor with L. G. Gorostiza, *Advances in Filtering and Optimal Stochastic Control*, Lecture Notes in Control and Inform. Sci., No. 42, Springer-Verlag, New York, Berlin, 1982.
- co-editor with I. Capuzzo-Dolcetta and T. Zolezzi, *Recent Advances in Dynamic Programming*, Lecture Notes in Math., No. 1119, Springer-Verlag, New York, Berlin, 1985.
- Controlled Markov Processes and Nonlinear Evolution Equations*, Accademia Nazionale dei Lincei, Scuola Normale Superiore, 1987.
- co-editor with P.-L. Lions, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA Vol. Math. Appl., No. 10, Springer-Verlag, New York, Berlin, 1987.
- with H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, Berlin, 1992.

RESEARCH PUBLICATIONS

- with L. C. Young, *A generalized notion of boundary*, Trans. Amer. Math. Soc., 76 (1954), pp. 457–484.
- On a class of games over function space and related variational problems*, Ann. of Math., 60 (1954), pp. 578–594.
- with R. Bellman and D. V. Widder, *Variational problems with constraints*, Ann. Mat. (4), 41 (1956), pp. 310–323.
- A note on differential games of prescribed duration*, Contrib. Theory Games, Ann. of Math. Stud., No. 36, pp. 407–412.
- with L. D. Berkovitz, *On differential games with an integral payoff*, Contrib. Theory Games, Ann. of Math. Stud., No. 39, pp. 413–435.
- An example in the problem of least area*, Proc. Amer. Math. Soc., 7 (1956), pp. 1063–1074.
- with L. C. Young, *Representations of generalized surfaces as mixtures*, Rend. Circ. Mat. Palermo (2), 5 (1956), pp. 117–144.
- with L. C. Young, *Generalized surfaces with prescribed elementary boundary*, Rend. Circ. Mat. Palermo (2), 5 (1956), pp. 320–340.
- Functions with generalized gradient and generalized surfaces*, Ann. Mat. (4), 46 (1957), pp. 93–104.
- Irreducible generalized surfaces*, Riv. Math. Univ. Parma (4), 8 (1957), pp. 251–281.
- Nondegenerate surfaces of finite topological type*, Trans. Amer. Math. Soc., 90 (1959), pp. 323–335.
- Nondegenerate surfaces and fine-cyclic surfaces*, Duke Math. J., 26 (1959), pp. 137–146.
- Functions whose partial derivatives are measures*, Illinois J. Math., 4 (1960), pp. 452–478.
- with R. Rishell, *An integral formula for total gradient variation*, Arch. Math. (Basel), 11 (1960), pp. 218–222.
- with H. Federer, *Normal and integral currents*, Ann. of Math., 72 (1960), pp. 458–520.
- The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1961), pp. 102–116.
- On the oriented plateau problems*, Rend. Circ. Mat. Palermo (2), 11 (1962), pp. 1–22.
- A Problem of Random Accelerations*, Rept. No. 403, Univ. of Wisconsin, Math. Research Center, Madison, WI, 1963.
- Some Markovian optimization problems*, J. Math. Mech., 12 (1963), pp. 131–140.
- The convergence problem for differential games II*, in Contributions to the Theory of Games, Princeton University Press, Princeton, NJ, 1964.
- The Cauchy problem for degenerate quasilinear parabolic equations*, J. Math. Mech., 13 (1964), pp. 987–1008.
- Flat chains over a finite coefficient group*, Trans. Amer. Math. Soc., 121 (1966), pp. 160–186.
- Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279; Erratum, J. Math. Anal. Appl., 19 (1966), p. 204.
- with M. Nisio, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.
- Stochastic Lagrange multipliers*, in Mathematical Theory of Control, Proc. Symp., Univ. of Southern California, 1967, Academic Press, New York, 1967, p. 443.
- Optimal control of partially observable diffusions*, SIAM J. Control, 6 (1968), pp. 194–214.
- Optimal continuous parameter stochastic control*, SIAM Review, 11 (1969), pp. 470–509.

- The Cauchy problem for a nonlinear first-order partial differential equation*, J. Differential Equations, 5 (1969), pp. 515–530.
- Controlled diffusions under polynomial growth conditions*, in Calculus of Variations and Control Theory, A. V. Balakrishnan, ed., Academic Press, New York, 1969, pp. 209–234.
- Stochastic control for small noise intensities*, SIAM J. Control, 9 (1971), pp. 473–517.
- Stochastically perturbed dynamical systems*, in Proc. Conf. on Stochastic Differential Equations, Edmonton, July, 1972; Rocky Mountain Math. J., 4 (1974), pp. 407–433.
- with C. H. Su, *Some one dimensional migration models in population genetics theory*, Theory Population Biol., 5 (1974), pp. 431–449.
- Diffusion processes in population biology*, Appl. Probab., 7 (1975), pp. 100–105.
- A selection-migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219–233.
- Generalized solutions in optimal stochastic control*, in Proc. Second Kingston Conf. on Differential Games, Marcel Dekker, 1977.
- Exit probabilities and optimal stochastic control*, Appl. Math. Optim., 4 (1978), pp. 329–346.
- Equilibrium distributions of continuous polygenic traits*, SIAM J. Appl. Math., 36 (1979), pp. 148–168.
- with M. Voit, *Some measure-valued Markov processes in population genetics theory*, Indiana Univ. Math. J., 28 (1979), pp. 817–844.
- Measure-valued processes in the control of partially observable stochastic systems*, Appl. Math. Optim., 6 (1980), pp. 271–285.
- with C.-P. Tsai, *Optimal exit probabilities and differential games*, Appl. Math. Optim., 7 (1981).
- with E. Pardoux, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.
- Nonlinear semigroup for controlled partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 286–301.
- with S. K. Mitter, *Optimal control and nonlinear filtering for nondegenerate diffusion processes*, Stochastics, 7 (1982), pp. 63–77.
- Logarithmic transformations and stochastic control*, in Advances in Filtering and Optimal Stochastic Control, Lecture Notes in Control and Inform. Sci., No. 42, Springer-Verlag, New York, Berlin, 1982.
- Stochastic calculus of variations and mechanics*, J. Optim. Theory Appl., 41 (1983), pp. 55–74.
- Optimal control of Markov processes*, in Proc. Internat. Congress of Mathematicians 1983 (Invited Plenary Address).
- with M. Nisio, *On stochastic relaxed controls for partially observed diffusions*, Osaka Math. J., 93 (1984), pp. 71–108.
- with S.-J. Sheu, *Stochastic variational formula for fundamental solutions of parabolic PDE*, Appl. Math. Optim., 13 (1985).
- A stochastic control approach to some large deviations problems*, in Proc. Conf. Recent Advances in Dynamic Programming, Rome, 1984; Lecture Notes in Math., No. 1119, Springer-Verlag, New York, Berlin, 1985, pp. 52–66.
- with P. E. Souganidis, *A PDE approach to asymptotic estimates for optimal exit probabilities*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1986), pp. 171–192.
- with P. E. Souganidis, *Asymptotic series and the method of vanishing viscosity*, Indiana Univ. Math. J., 35 (1986), pp. 425–447.
- with S. P. Sethi and H. M. Soner, *An optimal stochastic production planning problem with randomly fluctuating demand*, SIAM J. Control Optim., 25 (1987), pp. 1494–1502.
- with S.-J. Sheu and H. M. Soner, *On the existence of the dominant eigenvalue and its application to the large deviation properties of an ergodic Markov process*, Stochastics, 22 (1987), pp. 187–199.
- with D. Vermes, *Convex duality approach to the optimal control of diffusions*, SIAM J. Control Optim., 27 (1989), pp. 876–907.
- with P. E. Souganidis, *Value functions for two-player, zero-sum, stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–312.
- with H. M. Soner, *Asymptotic expansions for Markov processes with Levy generators*, Appl. Math. Optim., 19 (1989), pp. 203–223.
- Generalized solutions and convex duality in optimal control*, in Partial Differential Equations and the Calculus of Variations, F. Colombini et al., eds., Birkhauser, Boston, 1989, pp. 461–472.
- with E. Pardoux, *Piecewise monotone filtering with small observation noise*, SIAM J. Control Optim., 27 (1989), pp. 1156–1181.
- with D. Ji, P. Salame, and Q. Zhang, *Piecewise monotone filtering in discrete time with small observation noise*, IEEE Trans. Automat. Control, 36 (1991), pp. 1181–1186.
- with T. Zariphopoulou, *An optimal investment/consumption model with borrowing*, Math. Oper. Res., 16 (1991), pp. 802–822.

- with B. Fitzpatrick, *Numerical method for an optimal investment / consumption problem*, Math. Oper. Res., 16 (1991), pp. 823–841.
- with Q. Zhang, *Nonlinear filtering with small observation noise: Piecewise monotone observations*, in Stochastic Analysis, E. Merzbach, A. Shwartz, and E. Mayer-Wolf, eds., Academic Press, New York, 1991.
- with Q. Zhang, *Piecewise monotone filtering with small observation noise*, in Proc. Joint US–France Workshop in Stochastic Analysis, I. Karatzas and D. Ocone, eds., 1991.
- with M. James, *Asymptotic series and exit time probabilities*, Ann. Probab., to appear.
- with S. G. Grossman, J.-L. Vila, and T. Zariphopoulou, *Optimal portfolio rebalancing with transaction costs*, Econometrica, submitted.
- with W. McEneaney, *Risk Sensitive Optimal Control and Differential Games*, Brown Univ. LCDS, Rept. 92-1.

INVITED SUMMARY PAPERS PUBLISHED IN CONFERENCE PROCEEDINGS

- Optimal control of diffusion processes*, in Functional Analysis and Optimization, E. R. Cannello, ed., Academic Press, New York, 1966, pp. 68–84.
- Some problems of optimal stochastic control*, in Stochastic Optimization and Control, H. F. Karreman, ed., John Wiley, New York, 1968, pp. 59–64.
- Optimal continuous parameter stochastic control*, in Actes de Congress Internat. des Math., 1970; Gauthier-Villars, Paris, 1971, pp. 163–167.
- Nonlinear partial differential equations: Probabilistic and game theoretic methods*, Proc. CIME Summer School “Problems in Nonlinear Analysis”, Varenna, 1970, pp. 95–128.
- Dynamical systems with small stochastic terms*, in Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972.
- Optimal control of diffusion processes*, in Stochastic Differential Equations, J. B. Keller and H. P. McKean, eds., SIMA-AMS Proc., Vol. VI, 1973, pp. 163–171.
- Distributed parameter systems in population biology*, in Control Theory, Numerical Methods and Computer Systems Modelling, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Econ. and Math. Sys., No. 107, Springer-Verlag, New York, Berlin, 1975.
- with C. P. Tsai, *Some stochastic systems depending on small parameters*, Proc. Internat. Symp. on Dynamical Systems at Brown University, Academic Press, New York, 1976, pp. 103–114.
- Inclusion probability and optimal stochastic control*, IRIA Seminars Review, 1977.
- with C. P. Tsai, *Optimal inclusion probability and differential games*, IRIA Seminars Review, 1977.
- with M. Viot, *Some measure-valued population processes*, in Stochastic Analysis, A. Friedman and M. Pinsky, eds., Academic Press, New York, 1978, pp. 97–108.
- Optimal control of Markov diffusion processes*, Proc. Joint Automat. Control Conf., Vol. 1, 1987, pp. 355–358.
- Large deviations for diffusions depending on small parameters: A stochastic control method*, Proc. 1st AFCET-SMF Symp., Ecole Polytechnique, 1978.
- with C. P. Tsai, *Minimum exit probabilities and differential games*, Proc. 3rd Kingston Conf. on Differential Games and Control Theory, Marcel Dekker, 1978.
- with E. Pardoux, *Partially observed stochastic control systems*, Proc. 18th IEEE Conf. on Decision and Control, 1979, pp. 163–165.
- Lecture Notes on Diffusion Approximation and Optimal Stochastic Control*, Clemson University, 1979. *Stochastic control under partial observation*, 4th Internat. Conf. on Analysis and Optimization of Systems, INRIA, France, 1980.
- with O. Hernandez-Lerma, *Control optimo de procesos de diffusion Markovians*, Conferencias sobre sistemas estocasticas, Ciencia, 32 (1981), pp. 39–55.
- with S. K. Mitter, *Optimal control and nonlinear filtering of nondegenerate diffusions*, Proc. 20th IEEE Conf. on Decision and Control, 1981.
- with R. W. McGwier, *A regular perturbation expansion in nonlinear filtering*, Proc. 22nd IEEE Conf. on Decision and Control, 1983, pp. 82–83.
- with P. E. Souganidis, *A PDE approach to asymptotic estimates for optimal exit problems*, Lecture Notes in Control and Inform. Sci., Proc. IFIP Conf., Springer-Verlag, New York, Berlin, 1984.
- with P. E. Souganidis, *Asymptotic series for solutions to the dynamic programming equation for diffusions with small noise*, Proc. 24th IEEE Conf. on Decision and Control, Vol. 1, 1985.
- with H. M. Soner, *A stochastic production planning problem with random demand*, Proc. 24th IEEE Conf. on Decision and Control, Vol. 1, 1985.
- with D. Vermes, *Generalized solutions in the optimal control of diffusions*, Proc. IMA Workshop, IMA Vol. Math. Appl., No. 10, Springer-Verlag, New York, Berlin, 1987.
- with P. E. Souganidis, *Two-player, zero-sum stochastic differential games*, Analyse Mathematique et Applications, Gauthier-Villars, Paris, 1988.

- with D. Ji and E. Pardoux, *Piecewise linear filtering with small observations noise*, Proc. 8th INRIA Conf. on Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci., No. 111, Springer-Verlag, New York, Berlin, 1988.
- with E. Pardoux, *Piecewise monotone filtering with small observation noise*, Proc. 27 IEEE Conf. on Decision and Control, 1988.
- with B. Fitzpatrick, *Numerical methods for optimal investment-consumption model*, Proc. 29th IEEE Conf. on Decision and Control, 1990.

DISCRETE-TIME CONTROLLED MARKOV PROCESSES WITH AVERAGE COST CRITERION: A SURVEY*

ARISTOTLE ARAPOSTATHIS[†], VIVEK S. BORKAR[‡], EMMANUEL
FERNÁNDEZ-GAUCHERAND[§], MRINAL K. GHOSH¹, AND STEVEN I. MARCUS²

This paper is dedicated to Wendell Fleming on the occasion of his 65th birthday.

Abstract. This work is a survey of the average cost control problem for discrete-time Markov processes. The authors have attempted to put together a comprehensive account of the considerable research on this problem over the past three decades. The exposition ranges from finite to Borel state and action spaces and includes a variety of methodologies to find and characterize optimal policies. The authors have included a brief historical perspective of the research efforts in this area and have compiled a substantial yet not exhaustive bibliography. The authors have also identified several important questions that are still open to investigation.

Key words. controlled Markov processes, average cost, stationary policies, dynamic programming, optimal policies, ergodicity

AMS(MOS) subject classifications. 93E20, 60J70

1. Introduction. The average cost criterion (equivalently, the long-run average or ergodic cost) is a popular criterion for optimization of stochastic dynamical systems over an infinite time horizon. It is a reasonable criterion to use when the anticipated time interval for optimization (which in practice is finite) is long compared to other timescales involved, and there are no compelling reasons to prefer short-term optimization over long-term. Naturally, it is not favored in financial applications where money spent now is worth more than money spent later, but there are situations (communication networks being a prime example) where a “steady state” operation is expected over intervals that are long compared to the time constants of the system. Then it makes sense to minimize the limiting time-averaged cost, i.e., the “average cost.”

Mathematically, the criterion stands out as being much more difficult to analyze than the others; while other classical criteria lead to reasonably complete solutions, the average cost does not. The finite state and action problem is well understood, but there are numerous counterexamples in which infinite state or action problems do not have a nice solution. In fact, it appears not as a single problem but a collection of problems, some of which do not have a nice solution (cf. [150]). Thus, a variety of approaches have been developed to handle different situations. Not surprisingly,

*Received by the editors October 25, 1991; accepted for publication (in revised form) July 21, 1992. This work was supported in part by Texas Advanced Research Program (Advanced Technology Program) grants 003658-093 and 003658-186; in part by Air Force Office of Scientific Research grants AFOSR-91-0033, F49620-92-J-0045, and F49620-92-J-0083; and in part by National Science Foundation grant CDR-8803012.

[†]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712.

[‡]Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India.

[§]Systems and Industrial Engineering Department, University of Arizona, Tucson, Arizona 85721.

¹Department of Mathematics, Indian Institute of Science, Bangalore 560012, India.

²Department of Electrical Engineering and Systems Research Center, University of Maryland, College Park, Maryland 20742.

this is one chapter of Markov decision theory that is anything but closed. At the same time, it has come of age, having been studied for over 30 years, with promises of significant advances on the horizon. This, in short, is the *raison d'être* for this survey; we have attempted to put together a coherent account of what has been done, with an indication of what future advances may be.

Any such project has obvious limitations. Space constraints dictate a certain amount of selection, and not every relevant work can be covered in significant detail. We have included proofs where we felt they were essential to understanding the results or contained potentially useful novel ideas. In all cases, a serious attempt at objectivity has been made. For complementary reading on the general subject of Markov decision theory, see [137], [181], [196], [207].

The paper is organized as follows: §2 describes the problem formulation in full detail. Section 3 gives a brief history. Sections 4–6 extensively treat the finite state, the countable state, and the Borel state space cases, respectively, under complete observations. Section 7 treats the problem under partial observations. Section 8 describes some recent results on multiobjective average cost control. Finally, we conclude with some relevant remarks.

2. Preliminaries and formulation of the problem. In this section, the model and basic results concerning controlled Markov processes are given in the most general form needed for our presentation. In some subsequent sections, we specialize our presentation to situations in which measure-theoretic aspects are of no essential concern, as in the case for models with countable state space, allowing for a more transparent exposition. Before presenting the model, we summarize our key notation as follows:

- \mathbb{R} : set of real numbers;
- \mathbb{N} : set of positive integers;
- \mathbb{N}_0 : set of nonnegative integers;
- $\mathcal{B}(\mathbf{W})$: Borel σ -algebra of a given topological space \mathbf{W} ;
- $\mathcal{P}(\mathbf{W})$: for a Borel space \mathbf{W} (see [15], [82]), the set of all probability measures on $\mathcal{B}(\mathbf{W})$ endowed with the topology of weak convergence (see [134]).

The following are function spaces on a topological space \mathbf{W} :

- $C_b(\mathbf{W}) := \{v : \mathbf{W} \rightarrow \mathbb{R} \mid v \text{ is continuous and bounded}\}$;
- $\mathcal{M}(\mathbf{W}) := \{v : \mathbf{W} \rightarrow \mathbb{R} \mid v \text{ is Borel measurable}\}$;
- $\mathcal{M}_b(\mathbf{W}) := \{v : \mathbf{W} \rightarrow \mathbb{R} \mid v \text{ is Borel measurable and bounded}\}$;
- $\mathcal{L}(\mathbf{W}) := \{v : \mathbf{W} \rightarrow \mathbb{R} \mid v \text{ is lower semicontinuous and bounded below}\}$;
- $\mathcal{L}_b(\mathbf{W}) := \mathcal{L}(\mathbf{W}) \cap \mathcal{M}_b(\mathbf{W})$.

For $v \in \mathcal{M}_b(\mathbf{W})$, we let

- $\|v\| := \sup_{w \in \mathbf{W}} \{|v(w)|\}$;
- $\text{span}(v) := \sup_{w, w' \in \mathbf{W}} \{v(w) - v(w')\}$;
- $v^+ := v - \inf_{w \in \mathbf{W}} \{v(w)\}$, $v^- := v - \sup_{w \in \mathbf{W}} \{v(w)\}$.

We refer to $\text{span}(v)$ as the *span seminorm* of v .

The following is a list of the abbreviations used in this paper (the section where each abbreviation is first introduced is indicated in parenthesis):

- AC average cost (§2.4);
- ACOE average cost optimality equation (§3);
- ACOI average cost optimality inequality (§5.2);
- CMP controlled Markov process (§2.1);

- CO completely observable (§3);
- DC discounted cost (§2.4);
- DCOE discounted cost optimality equation (§2.6);
- PO partially observable (§7.2);
- POCMP partially observable controlled Markov process (§3);
- TC total cost (§2.4).

2.1. The model. A discrete-time, stationary controlled Markov process (CMP), or Markov decision process, is a stochastic dynamical system specified by the five-tuple $(\mathbf{S}, \mathbf{A}, U, P, c)$, where

- (a) \mathbf{S} is a Borel space, called the *state space*, the elements of which are called *states*;
- (b) \mathbf{A} is a Borel space, called the *action* or *control* space;
- (c) $U : \mathbf{S} \rightarrow \mathcal{B}(\mathbf{A})$ is a strict, measurable, compact-valued multifunction (see the Appendix). $U(x)$ represents the set of admissible actions (or control inputs) when the system is in state $x \in \mathbf{S}$. Accordingly, the set of admissible state/action pairs is $\mathbf{K} := \{(x, a) : x \in \mathbf{S}, a \in U(x)\} = \text{Graph}(U)$, and we have that $\mathbf{K} \in \mathcal{B}(\mathbf{S} \times \mathbf{A})$. This set is endowed with the subspace topology corresponding to $\mathcal{B}(\mathbf{S} \times \mathbf{A})$;
- (d) P is a stochastic kernel on \mathbf{S} given \mathbf{K} , called the *transition kernel*. It is assumed to be Borel measurable, i.e., $P(D \mid \cdot) : \mathbf{K} \rightarrow [0, 1]$ is Borel measurable, for each $D \in \mathcal{B}(\mathbf{S})$;
- (e) $c : \mathbf{K} \rightarrow \mathbb{R}$ is the (measurable) one-stage cost function.

The evolution of the system is as follows. Let X_t denote the state at time $t \in \mathbb{N}_0$, and A_t the action chosen at that time. If $X_t = x \in \mathbf{S}$ and $A_t = a \in U(x)$, then (i) a cost $c(x, a)$ is incurred, and (ii) the system moves to the next state X_{t+1} , according to a probability distribution $P(\cdot \mid x, a)$. Once the transition into the next state has occurred, a new action is chosen, and the process is repeated.

The total period of time over which the system is to be observed is called the planning (or decision-making or control) horizon and is denoted by T . It can be a finite interval $\{0, \dots, N-1\}$, with $N \in \mathbb{N}$, or an infinite horizon, e.g., $T = \mathbb{N}_0$.

The (admissible) *history spaces* are defined as

$$\mathbf{H}_0 := \mathbf{S}, \quad \mathbf{H}_t := \mathbf{H}_{t-1} \times \mathbf{K}, \quad t \in \mathbb{N}_0,$$

and the canonical sample space is defined as $\mathbf{\Omega} := (\mathbf{S} \times \mathbf{A})^\infty$. These spaces are endowed with their respective product topologies and are therefore Borel spaces. A generic element $\omega \in \mathbf{\Omega}$ is of the form $\omega = (x_0, a_0, x_1, a_1, \dots)$, $x_i \in \mathbf{S}$, $a_i \in \mathbf{A}$; all random variables will be defined on the measurable space $(\mathbf{\Omega}, \mathcal{B}(\mathbf{\Omega}))$.

The state, action (or control), and information processes, denoted by $\{X_t\}_{t \in T}$, $\{A_t\}_{t \in T}$ and $\{H_t\}_{t \in T}$, respectively, are defined by the projections

$$X_t(\omega) := x_t, \quad A_t(\omega) := a_t, \quad H_t(\omega) := (x_0, \dots, a_{t-1}, x_t), \quad t \in T$$

for each realization $\omega = (x_0, \dots, a_{t-1}, x_t, a_t, \dots) \in \mathbf{\Omega}$. Since $\mathcal{B}(\mathbf{\Omega}) = (\mathcal{B}(\mathbf{S}) \times \mathcal{B}(\mathbf{A}))^\infty$, the above are well-defined random processes on $(\mathbf{\Omega}, \mathcal{B}(\mathbf{\Omega}))$. Note that $\mathcal{B}(\mathbf{\Omega}) = \bigvee_{t=0}^\infty \mathfrak{F}_t$, where $\mathfrak{F}_t = \sigma(H_t)$, the σ -algebra generated by H_t .

Example 2.1. Let \mathbf{S} , \mathbf{A} , \mathbf{W} be Borel spaces and $F : \mathbf{S} \times \mathbf{A} \times \mathbf{W} \rightarrow \mathbf{S}$ a Borel function. Consider a nonlinear stochastic system described by the system equation

$$X_{t+1} = F(X_t, A_t, W_t), \quad t \in T,$$

where the process $\{W_t\}$ is a sequence of independent and identically distributed (i.i.d.) \mathbf{W} -valued random variables, with common probability distribution \mathcal{P}_W , often referred to as a stochastic state disturbance, or noise; $\{W_t\}$ is assumed to be independent of X_0 . Suppose that a strict, measurable, compact-valued multifunction $U : \mathbf{S} \rightarrow \mathcal{B}(\mathbf{A})$ has been specified, giving the necessary constraints on the control actions, or that $U(x) = \mathbf{A}$, for all $x \in \mathbf{S}$, if there are no constraints. Then the evolution of the system is equivalently described in terms of the stochastic kernel P on \mathbf{S} given \mathbf{K} defined as

$$P(D \mid x, a) := \int_{\mathbf{W}} I\{F(x, a, w) \in D\} \mathcal{P}_W(dw), \quad (x, a) \in \mathbf{K}, \quad D \in \mathcal{B}(\mathbf{S}),$$

where $I\{A\}$ denotes the indicator function of the event A . The additional specification of a measurable cost function $c : \mathbf{K} \rightarrow \mathbb{R}$ would completely define a CMP $(\mathbf{S}, \mathbf{A}, U, P, c)$.

Example 2.2. Consider a countable set \mathbf{S} endowed with the discrete topology. With no loss in generality we can take $\mathbf{S} = \mathbb{N}_0$. Let \mathbf{A} be a Borel space and $U(x) = \mathbf{A}$, for all $x \in \mathbf{S}$. In this case, every stochastic kernel on \mathbb{N}_0 given $\mathbf{K} := \mathbb{N}_0 \times \mathbf{A}$ reduces to a collection of discrete probability distributions parameterized by $(i, a) \in \mathbf{K}$. These can also be represented by a collection of stochastic matrices $\{P(a) = [p_{ij}(a)]\}_{a \in \mathbf{A}}$; i.e., $P(a)$ is a state transition matrix, and $p_{ij}(a)$ is the probability that the state of the system makes a transition from i to j , under action a . Therefore, additionally specifying a cost function $c : \mathbb{N}_0 \times \mathbf{A} \rightarrow \mathbb{R}$ completely defines a CMP.

2.2. Policies and performance criteria. An *admissible control strategy*, or *policy*, is a sequence $\pi = \{\pi_t\}_{t \in T}$ of Borel measurable stochastic kernels on \mathbf{A} given \mathbf{H}_t , satisfying the constraint

$$\pi_t(U(x_t) \mid h_t) = 1, \quad x_t \in \mathbf{S}, \quad h_t \in \mathbf{H}_t.$$

The set of all admissible policies will be denoted by Π .

If $\mu \in \mathcal{P}(\mathbf{S})$ and $\pi \in \Pi$ are given, there exists a unique probability measure \mathcal{P}_μ^π on $(\Omega, \mathcal{B}(\Omega))$ satisfying the following [15, Prop. 7.28, pp. 140–144], [130, Prop. V.1.1, pp. 162–164], with $D \in \mathcal{B}(\mathbf{S})$ and $C \in \mathcal{B}(\mathbf{A})$:

$$(2.1) \quad \mathcal{P}_\mu^\pi(X_0 \in D) = \mu(D),$$

$$(2.2) \quad \mathcal{P}_\mu^\pi(A_t \in C \mid H_t) = \pi_t(C \mid H_t), \quad \mathcal{P}_\mu^\pi\text{-a.s.},$$

$$(2.3) \quad \mathcal{P}_\mu^\pi(X_{t+1} \in D \mid H_t, A_t) = P(D \mid X_t, A_t), \quad \mathcal{P}_\mu^\pi\text{-a.s.}$$

Therefore, if μ is the distribution of the initial state X_0 , and policy $\pi \in \Pi$ is used, the underlying probability space of all random variables of interest is $(\Omega, \mathcal{B}(\Omega), \mathcal{P}_\mu^\pi)$. The expectation operator with respect to \mathcal{P}_μ^π will be denoted by E_μ^π . Furthermore, if μ is a Dirac measure at $x \in \mathbf{S}$, we will simply write \mathcal{P}_x^π and E_x^π .

Certain classes of admissible policies are of special interest. A policy π is called a *Markov randomized policy* if there exists a sequence of measurable maps $\{f_t\}_{t \in T}$, called *randomized decision rules*, where $f_t : \mathbf{S} \rightarrow \mathcal{P}(\mathbf{A})$, for each $t \in T$, such that

$$\pi_t(\cdot \mid H_t) = f_t(X_t)(\cdot), \quad \mathcal{P}_\mu^\pi\text{-a.s.}$$

Conversely, every sequence of measurable maps $f_t : \mathbf{S} \rightarrow \mathcal{P}(\mathbf{A})$, $t \in T$, satisfying $f_t(x)(U(x)) = 1$, defines a Markov randomized policy in an obvious way; with some

abuse in notation, the sequence itself will be referred to as the policy. The set of all Markov randomized policies will be denoted by Π_M . A policy $\{f_t\}_{t \in T} \in \Pi_M$ is called a *stationary* randomized policy if there is a randomized decision rule f such that, for all $t \in T$, $f_t = f$. The set of all stationary randomized policies will be denoted by Π_{SR} . A *nonrandomized*, *deterministic*, or *pure* decision rule is a measurable map $f: \mathcal{S} \rightarrow \mathcal{A}$. A policy $\{f_t\}_{t \in T} \in \Pi_M$ is called a *nonrandomized*, *deterministic*, or *pure* Markov policy if each f_t is deterministic. Hence, in this case, $A_t = f_t(X_t)$ almost surely. The set of deterministic Markov policies will be denoted by Π_{MD} . Stationary deterministic policies are defined in the obvious way. The set of all stationary deterministic policies is denoted by Π_{SD} , and, for $\pi \in \Pi_{SD}$, $\pi(x)$ will denote the action chosen at $x \in \mathcal{S}$. Clearly $\Pi_{SD} \subseteq \Pi_{MD} \subseteq \Pi_M \subseteq \Pi$, and $\Pi_{SD} \subseteq \Pi_{SR} \subseteq \Pi_M$.

It is easily seen that, under a policy $\pi = \{f_t\}_{t \in T} \in \Pi_M$, the state process $\{X_t\}_{t \in T}$ is a Markov process. That is, for $D \in \mathcal{B}(\mathcal{S})$,

$$\begin{aligned} \mathcal{P}_\mu^\pi(X_{t+1} \in D \mid X_t, \dots, X_0) &= \mathcal{P}_\mu^\pi(X_{t+1} \in D \mid X_t) \\ &= \int_{\mathcal{A}} P(D \mid X_t, a) f_t(X_t)(da), \quad \mathcal{P}_\mu^\pi\text{-a.s.}, \end{aligned}$$

and, under a policy $\pi' \in \Pi_{SR}$, $\{X_t\}_{t \in T}$ is a Markov process with stationary transition probabilities.

Each policy $\pi \in \Pi$ incurs a stream of random costs, e.g., $\{c(X_t, f_t(X_t))\}_{t \in T}$, for $\{f_t\}_{t \in T} \in \Pi_{MD}$. Depending upon the problem requirements, several cost evaluation criteria are studied. The following criteria are frequently used.

Total cost (TC). The total cost incurred by the policy $\pi \in \Pi$ over the entire planning horizon is given by

$$J_T(\mu, \pi) := E_\mu^\pi \left[\sum_{t \in T} c(X_t, A_t) \right].$$

When the horizon is finite, i.e., $T = \{0, \dots, N-1\}$, $N \in \mathbb{N}_0$, we denote the above more explicitly as $J_N(\mu, \pi)$. Furthermore, given a *terminal cost* function $h \in \mathcal{M}_b(\mathcal{S})$, we define

$$J_N(\mu, \pi, h) := E_\mu^\pi \left[\sum_{t=0}^{N-1} c(X_t, A_t) + h(X_N) \right].$$

Discounted cost (DC). Let $0 < \beta < 1$, the *discount factor*, and $\pi \in \Pi$ be given. The total discounted cost incurred by π over the infinite planning horizon is given by

$$J_\beta(\mu, \pi) := E_\mu^\pi \left[\sum_{t=0}^{\infty} \beta^t c(X_t, A_t) \right].$$

Average cost (AC). The expected long-run average cost incurred by $\pi \in \Pi$ is given by

$$J(\mu, \pi) := \limsup_{N \rightarrow \infty} E_\mu^\pi \left[\frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t) \right] = \limsup_{N \rightarrow \infty} \frac{1}{N} J_N(\mu, \pi).$$

Sample path average cost. This is a pathwise version of the AC, and, for $X_0 = x$, it is given by

$$J_S(x, \pi) := \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t),$$

where $\{X_t\}$ and $\{A_t\}$ are the state and control process induced by $\pi \in \Pi$. Here, $J_S(x, \pi)$ is to be regarded as an extended real-valued random variable on the canonical sample space.

For the AC criterion, the limit of the expected average cost may not exist for some or all policies $\pi \in \Pi$, and thus the limit superior is used. This is always well defined and captures the *worst* possible asymptotic expected average performance under policy $\pi \in \Pi$; i.e., it gives a “pessimistic” measure of performance. On the other hand, the limit inferior could also be used, which would yield an “optimistic” measure of performance by capturing the *best* possible asymptotic expected average performance. The planning horizon for the TC criterion can be finite or infinite, whereas, for the other criteria above, it is always infinite. Under certain conditions, it can be shown that a problem with the DC criterion is equivalent to one with a TC criterion, with a random (finite) horizon; see [40, pp. 31–32]. Also, it can be shown that, for each $\pi \in \Pi$, a policy $\pi' \in \Pi_M$ can be found such that $E_\mu^\pi[c(X_t, A_t)] = E_\mu^{\pi'}[c(X_t, A_t)]$, for each $t \in \mathbb{N}_0$ and any initial distribution $\mu \in \mathcal{P}(\mathcal{S})$ [42], [51, §3.8]. Thus, for criteria that are determined by these expected costs, such as the AC, DC, and TC criteria, it suffices to consider policies in Π_M .

For an infinite planning horizon, $J_T(\mu, \pi)$ need not be well defined or may be infinite for all $\pi \in \Pi$, rendering this criterion useless for comparing the performance under different policies. Therefore, the DC or AC criteria are usually selected when the planning horizon is infinite. When the DC criterion is used, a rather complete theory is available for the corresponding dynamic programming formulation of the problem [14], [15], [51], [82], [103], [150], [200]. In this situation, future costs are discounted at a fixed rate $0 < \beta < 1$, and therefore, if β is not sufficiently close to 1, the asymptotic behavior of the state/cost process may not be important at all. Quite the opposite is the case with the AC criterion, under which all decision times are given equal weight, and we take the limit of time-averaged expected costs. The finite time evolution of the state/cost process is, in some sense, completely irrelevant in this case, and some sort of asymptotic stable behavior is desired, making this case mathematically much more involved than the previous one. Hence, the DC and AC can be seen as two opposite extremes in the spectrum of possible criteria that can be considered, in the sense that the first one captures primarily the performance of the process at the present and near future, and the second captures the performance at the distant future.

2.3. The optimal control problem. The *optimal control (or decision) problem* is that of selecting an admissible policy such that a given performance criterion is minimized over all admissible policies. For example, for the DC criterion, a policy $\pi^* \in \Pi$ is said to be (β) -discount ε -optimal for the initial distribution μ if

$$J_\beta(\mu, \pi^*) \leq J_\beta(\mu, \pi) + \varepsilon \quad \forall \pi \in \Pi,$$

where $\varepsilon > 0$. If a policy is discount ε -optimal for all distributions $\mu \in \mathcal{P}(\mathcal{S})$, then it is simply called discount ε -optimal. If a policy is discount ε -optimal *for all* $\varepsilon > 0$, then it is called *discount optimal*. The (optimal) value function is given by

$$(2.4) \quad J_\beta^*(\mu) := \inf_{\pi \in \Pi} J_\beta(\mu, \pi).$$

Also, if μ is concentrated at $x \in \mathcal{S}$, we denote the value function by $J_\beta^*(x)$. Similar definitions apply to other criteria; $J_T^*(\mu)$ and $J^*(\mu)$ will denote the optimal value functions for the TC and AC criteria, respectively. For sample path AC, we define an

optimal policy as follows: We say that a policy $\pi^* \in \Pi$ is *sample path AC optimal* (or *almost surely AC optimal*) if there exists a constant ρ^* such that, for any initial law μ ,

$$J_S^*(\mu, \pi^*) = \rho^*, \quad \mathcal{P}_\mu^{\pi^*}\text{-a.s.},$$

while, for any other policy $\pi \in \Pi$ and any initial law μ' ,

$$J_S^*(\mu', \pi) \geq \rho^*, \quad \mathcal{P}_{\mu'}^\pi\text{-a.s.}$$

The constant ρ^* is the sample path optimal average cost.

Having defined various optimality criteria and the set of admissible policies Π , the obvious question now is: Do there exist optimal policies? Without imposing further assumptions on our general model, the answer is no. One of the reasons behind this is that the Borel measurability assumption in the definition of admissible policies is too restrictive, in general, to be able to attain the infimum in (2.4). To circumvent this problem, either a broader sense of measurability is allowed, i.e., a larger set of admissible policies is used, or further assumptions are imposed. The first approach was taken by Shreve and Bertsekas [15], [164], [165], who considered *universally measurable* policies, a class properly containing the (Borel measurable) admissible policies defined previously; see also [51]. We will instead follow the second approach mentioned above and concentrate on the *semicontinuous model*, as studied in [15], [47], [51], [71], [88], [123], [152]–[154].

2.4. The semicontinuous model. In general, we consider the case when the one-stage cost function $c(\cdot, \cdot)$ is unbounded. Since, for the most part, the criteria considered in this paper are given by a sum of expected costs over the infinite horizon, then, to avoid indeterminate situations, the following conditions will be assumed to hold throughout the paper, unless otherwise indicated.

Assumption 2.1. $c(x, a) \geq 0$ for all $(x, a) \in \mathbf{K}$.

Assumption 2.2. The transition kernel $P(\cdot \mid x, a)$ is *weakly continuous* in (x, a) ; that is, $v(\cdot) \in C_b(\mathbf{S})$ implies that $\int_{\mathbf{S}} v(y)P(dy \mid \cdot, \cdot) \in C_b(\mathbf{K})$.

Assumption 2.3. (i) The multifunction $U(x)$ is upper semicontinuous; (ii) $c(\cdot, \cdot) \in \mathcal{L}(\mathbf{K})$.

Remark 2.1. Concerning Assumption 2.1, note that (for the AC and DC criteria) we must only assume that the cost is bounded below. The assumption that the cost is nonnegative is only made for convenience. Assumption 2.2 is equivalent to $\int v(y)P(dy \mid \cdot, \cdot) \in \mathcal{L}(\mathbf{K})$, for each $v(\cdot) \in \mathcal{L}(\mathbf{S})$ [51, p. 52]. This property is crucial in our development.

Example 2.3. For the nonlinear stochastic system in Example 2.1, assume further that

- (i) \mathbf{A} is compact,
- (ii) For each $x \in \mathbf{S}$, $U(x)$ is closed (and therefore compact), and
- (iii) The system function $F : \mathbf{K} \times \mathbf{W} \rightarrow \mathbf{S}$ is continuous.

If $c(\cdot, \cdot) \in \mathcal{L}(\mathbf{K})$, then, by Remark 2.1, Assumption 2.2 will hold. Furthermore, the assumption on the compactness of \mathbf{A} can be dispensed with if there are compact subsets $\mathbf{K}_1 \subseteq \mathbf{K}_2 \subseteq \cdots$ in $\mathbf{S} \times \mathbf{A}$, such that $\mathbf{K} = \bigcup_{n \in \mathbb{N}} \mathbf{K}_n$ and

$$\liminf_{n \rightarrow \infty} \{c(x, a) : (x, a) \in \mathbf{K}_n \setminus \mathbf{K}_{n-1}\} = +\infty,$$

since, in this case, \mathbf{A} can be conveniently compactified; cf. [15, Cor. 8.6.1, p. 210]. Also, the case in which $\mathbf{S} = \mathbb{R}^n$, $\mathbf{A} = \mathbb{R}^m$, and $c(x, a) = x'Qx + a'Ra$, where Q and

R are positive semidefinite and positive definite matrices, respectively, of appropriate dimensions can also be considered by a (one-point) compactification of \mathbf{A} [164, pp. 965–966].

Under Assumptions 2.1–2.3, the *undiscounted dynamic programming map* T given by

$$(2.5) \quad T(v)(x) := \inf_{a \in U(x)} \left\{ c(x, a) + \int_{\mathbf{S}} v(y) P(dy \mid x, a) \right\} \quad \forall x \in \mathbf{S}$$

maps $\mathcal{L}(\mathbf{S})$ into itself. Also, for $0 < \beta < 1$, the *discounted dynamic programming map* $T_\beta : \mathcal{L}(\mathbf{S}) \rightarrow \mathcal{L}(\mathbf{S})$ is given by

$$(2.6) \quad T_\beta(v) := T(\beta v).$$

The following properties are easily verified.

LEMMA 2.1. *Let $v, v' \in \mathcal{L}(\mathbf{S})$. Then (i) for all $k \in \mathbb{R}$, $T(v + k) = T(v) + k$; (ii) if $v \leq v'$, then $T(v) \leq T(v')$.*

Some key results for the stochastic control problem under a DC criterion are summarized in the following theorem.

THEOREM 2.1. *Under Assumptions 2.1–2.3*

(i) *The following equation, which is called the discounted cost optimality equation (DCOE), holds:*

$$(2.7) \quad J_\beta^*(x) = T_\beta(J_\beta^*)(x) = \inf_{a \in U(x)} \left\{ c(x, a) + \beta \int_{\mathbf{S}} J_\beta^*(y) P(dy \mid x, a) \right\}, \quad x \in \mathbf{S};$$

(ii) *A policy $\pi^* \in \Pi_{SD}$ is discount optimal if and only if $\pi^*(x)$ attains the infimum in (2.7), for all $x \in \mathbf{S}$;*

(iii) *A discount optimal policy $\pi^* \in \Pi_{SD}$ exists;*

(iv) *Define $T_\beta^0 : \mathcal{L}(\mathbf{S}) \rightarrow \mathcal{L}(\mathbf{S})$ as the identity operator and $T_\beta^k : \mathcal{L}(\mathbf{S}) \rightarrow \mathcal{L}(\mathbf{S})$, $k \in \mathbb{N}$, by $T_\beta^k(f) := T_\beta(T_\beta^{k-1}(f))$. Then, for any $f \in \mathcal{L}_b(\mathbf{S})$,*

$$T_\beta^k(f)(x) \xrightarrow[k \rightarrow \infty]{} J_\beta^*(x) \quad \text{for all } x \in \mathbf{S};$$

(v) *$J_\beta^*(\cdot)$ is nonnegative and lower semicontinuous.*

Remark 2.2. The above results are essentially contained in [15], [51]. The existence of a measurable selector that attains the infimum in (2.7), e.g., the result in (iii) of Theorem 2.1, follows from [15, Prop. 7.33, p. 153], [29], [47, pp. 35–38], [51, §2.6], [88], [139, Thm. 4.1, p. 9], [184, Thm. 9.1, p. 880]. The scheme used in (iv) of Theorem 2.1 to compute $J_\beta^*(\cdot)$ is called the *value iteration* (or successive approximations) algorithm. When the one-stage cost function is bounded, the usual approach is to prove the existence of a unique solution to the DCOE via a contraction mapping theorem [14], [82]. Otherwise, $J_\beta^*(\cdot)$ is not necessarily the only fixed point of T_β ; however, $J_\beta^*(\cdot)$ is the *minimal* fixed point of T_β among the class of nonnegative functions in $\mathcal{L}(\mathbf{S})$ [15, Chap. 5], [173].

3. A sketch of historical development. We now present a brief historical sketch of the development of CMP, with an emphasis on the average cost criterion. The roots of CMP can be traced back to the pioneering work of Wald [186], [187]

on sequential analysis and statistical decision functions. In the late 1940s and early 1950s, several investigators formulated the essential concepts of CMP, which are found in their work in sequential game models. A CMP can be viewed as a one-player game. Of particular interest is the work of Bellman and Blackwell [12], Bellman and LaSalle [13], and also Shapley, who formulated the essential mechanism of stochastic dynamic programming and used the theory of contraction mappings [160]. Using his famous heuristic “minimum cost to go,” Bellman showed how powerful the dynamic programming technique was by using it to solve problems in a myriad of settings [9]–[11]. Bellman studied mostly problems with a finite horizon, for which the backward induction approach of dynamic programming suffices to give a complete treatment. The situation is quite different in problems over an infinite horizon. Early work on CMP is also reported in econometrics [4], [49].

Howard [95] was apparently the first to study CMP with an average cost criterion. His *policy iteration* algorithm was the first major computational breakthrough, and his book helped establish CMP as an independent subject of investigation. For CMP with finite state and action spaces, Howard’s policy iteration scheme established the existence of a stationary deterministic policy, optimal in this class only. Derman [38] and Viskov and Shiryaev [183] independently showed that this policy was optimal among all admissible policies. Other computational methods were later proposed. Manne [125] gave a linear programming formulation for the AC criterion, and Wagner [185] later characterized extreme-point optima of the linear program as stationary deterministic policies. White [197] introduced the value iteration (or successive approximations) technique. Excellent accounts of these and other computational methods are given in [14, §5.2] and [137].

On the theoretical side, Blackwell’s seminal paper [18] gave considerable impetus to research in this area, motivating numerous other papers. In [18] Blackwell studied CMP with finite state and action spaces. He considered the DC criterion in great detail and established many important results. In the same paper, he initiated an approach for the AC case, which we will refer to as the *vanishing discount approach*: he treated the AC case as a limit of the DC case, when the discount factor goes to 1, i.e., the discounting effect vanishes. Blackwell established in [18] the existence of a stationary deterministic policy that is discount optimal, for all β sufficiently close to 1. This type of optimality is now called *Blackwell optimality* [14, pp. 336–341]. The relation between the discounted and average case also becomes apparent via Tauberian theorems [87, §4.6]. This fact seems to have been observed first by Gillette [79], who used Tauberian theorems to establish the existence of optimal stationary policies in a stochastic game problem with an AC criterion. Also, using Tauberian theorems, Derman [38] showed that the Blackwell optimal policy found in [18] was also optimal for the AC criterion. Average cost CMP with finite state and arbitrary action spaces were studied under various conditions in the works of [35], [57]–[59], [100].

Blackwell optimal policies do not necessarily exist when the state space is countably infinite [122]. In fact, average optimal policies need not exist in this situation [121], [150]. Similar nonexistence result holds when the state space is finite, but the action space is an arbitrary compact metric space [8]. For such models, the existence of an optimal policy has been proved by Bather [8], Martin-Löf [126], and Feinberg [58], under certain conditions. Derman [39] studied the problem with countable state space, finite action space, and bounded cost. He studied the following equation, which

became known as the *average cost optimality equation* (ACOE):

$$\rho + h(i) = \min_{a \in U(i)} \left\{ c(i, a) + \sum_{j \in S} P(j \mid i, a) h(j) \right\},$$

where ρ is a scalar, $h : S \rightarrow \mathbb{R}$, $S = \mathbb{N}_0$, and we write $P(j \mid \cdot, \cdot)$ for $P(\{j\} \mid \cdot, \cdot)$. He showed that, if the ACOE has a *bounded solution*, i.e., a solution (ρ, h) with $h(\cdot)$ a bounded function, then the stationary deterministic policy realizing the pointwise minimum on the right-hand side of the ACOE is average optimal, and ρ is the minimum average cost. Derman's paper, in conjunction with Derman and Veinott [43], showed that a sufficient condition for the existence of such a solution was that the expected hitting time of a fixed state under *any* stationary deterministic policy is bounded uniformly with respect to the choice of the policy and the initial state. Motivated by Blackwell's work, Taylor [177] extended the vanishing discount approach to obtain a bounded solution for a Markovian sequential replacement problem by studying the asymptotics of the *differential* discounted value function $h_\beta(\cdot) := J_\beta(\cdot) - J_\beta(0)$. Ross [147], [148] refined Taylor's procedure and showed that, under the Derman–Veinott [43] condition, $\{h_\beta(\cdot)\}_{\beta \in (0,1)}$ was uniformly bounded in β . By letting $\beta \uparrow 1$, Ross established that the ACOE had a bounded solution. This made the vanishing discount approach very popular. In subsequent works, many variants of Derman–Veinott recurrence conditions appeared. See [52], [178] for a great variety of such conditions. These conditions are difficult to remove, and counterexamples abound [150]. Actually, it has been shown in [64], in a very general setting, that the uniform boundedness of $\{h_\beta(\cdot)\}_{\beta \in (0,1)}$ in β is also a *necessary* condition for a bounded solution to the ACOE to exist. Cavazos-Cadena [30], [31], under some additional conditions, showed that the existence of bounded solutions to the ACOE necessarily impose a very strong recurrence structure on the model. Lippman [115] studied controlled semi-Markov processes with unbounded cost with both discounted and average cost criteria. Following the vanishing discount approach, he derived results for the average cost case under several restrictive assumptions. Federgruen, Hordijk, and Tijms [53] have explored the same approach.

Hordijk [91] extended many earlier results to countable state space and compact action spaces. He introduced the Lyapunov function method for CMP. He used this method to obtain a (possibly *unbounded*) solution to the ACOE, yielding an optimal policy. However, the Lyapunov function method necessarily imposes a blanket stability of the processes (in the sense of positive recurrence). Such stability is not always met in, e.g., many queueing model applications. In addition, he introduced some new concepts, particularly based on the relation of stochastic dynamic programming with Markov potential theory. There is a vast amount of literature devoted to CMP in several volumes of the Mathematisch Centrum tracts; see [181] and the references therein.

With Hordijk's work, it appeared that a shift away from the vanishing discount approach was necessary. Rosberg, Varaiya, and Walrand [144] treated the average cost criterion as the limiting case of the finite horizon problem, but details of their arguments depend heavily on the specifics of the problem they consider, viz., the control of two queues in tandem with a linear cost structure. Federgruen and Tijms [56] initiated a direct study of the ACOE by a span seminorm method, for bounded costs. This method allows us to obtain useful value iteration algorithms. Later, Federgruen, Schweitzer, and Tijms [55] treated the problem with countable state space

and unbounded costs. Assuming a recurrence condition on the model, they established the existence of a (possibly unbounded) solution to the ACOE, thereby establishing the existence of an optimal stationary deterministic policy.

In a series of papers [20]–[25], Borkar presented a convex analytic approach to treat the problem with countable state space, compact action space, and unbounded cost. This approach can be seen as an extension of the ideas in Manne [125] and Wagner [185]. Borkar stressed the existence of an optimal *stable* stationary deterministic policy, i.e., one that induces a positive recurrent process. While a blanket stability assumption (e.g., of Lyapunov type) may be too restrictive to cover many queueing applications, it nevertheless is desirable that the optimal policy be stable. Borkar showed that, to obtain an optimal stable stationary deterministic policy, either a blanket stability hypothesis or a condition on the cost that penalizes unstable behavior is necessary. He also emphasized the concept of almost sure optimality by a “pathwise” treatment of the problem. A comprehensive account of the convex analytic approach to CMP is given in [26].

After the extensive works of Hordijk, Federgruen et al., and Borkar, it seemed that the vanishing discount approach was not appropriate for many classes of problems with unbounded costs. However, this approach has been revived and generalized to a great extent in [17], [61], [63], [74], [76], [77], [83], [85], [155], [156], [167], [172], [190]. In some of these references, an *inequality* version of the ACOE is studied. In view of the results of [30], [31], and [64], it is clear that a bounded solution to the ACOE is too restrictive, in general. A natural candidate solution is one that is bounded below [28], [76], [85], [155], [156], [172], [190], or one having suitable growth properties [28] or satisfying other conditions [167]. Weber and Stidham [172], [190] studied the problem for queueing systems. Under a penalizing condition on the cost and some structural assumptions, they established the existence of a (possibly unbounded) solution to the ACOE and showed the existence of an optimal stationary deterministic policy. Sennott proceeded along similar lines. She identified very general conditions on the discounted value function so that the vanishing discount approach could successfully be pursued. We refer to [155]–[157], [172], [190] for many interesting examples of queueing systems and to [34] for a comparison of different sets of assumptions. Extensions of these techniques to semi-Markov decisions processes with applications to queueing systems have been reported in [157].

The first attempt to give a description of CMP with more general state and actions spaces was carried out by Karlin [98]. Blackwell [19], Maitra [123], and Strauch [173] studied CMP with a general state space and the discounted cost criterion. Their work was significantly extended by Shreve and Bertsekas in [15], [164], [165]. Feinberg [60] studied CMP with Borel state space and with arbitrary numerical criteria, which include TC, AC, and DC as particular cases. By establishing the convexity of the set of strategic measures (measures of the type \mathcal{P}_μ^π on the canonical space), he established the existence of an ε -optimal $f \in \Pi_{SD}$ for these criteria. De Leve [112]–[114] considered general state and action space CMP in continuous time with an AC criterion, with an emphasis on the ergodic behavior of the processes. Ross [148] used the vanishing discount approach to study CMP with an AC criterion, general state space, finite action space, and bounded cost function. He showed that, if the family of differential discounted value functions $\{h_\beta(\cdot)\}_{\beta \in (0,1)}$ is equicontinuous and uniformly bounded, then the ACOE admits a bounded solution, yielding an optimal stationary deterministic policy. Ross also introduced the concept of minorant. He showed that,

if there exists a state $x_0 \in \mathcal{S}$ and $\alpha > 0$ such that

$$P(x_0 \mid x, a) > \alpha \quad \text{for all } a \in U(x), \quad x \in \mathcal{S},$$

then the average cost case could be reduced to a discounted one. This was greatly extended in the work of Gubenko and Statland [80] (see also [43]). They showed that, under similar minorant (or majorant) conditions, a contraction map, with respect to the sup norm, could be defined on $\mathcal{M}_b(\mathcal{S})$, which would yield a bounded solution to the ACOE. They also obtained bounded solutions to the ACOE under continuity and boundedness conditions, which guarantee that $\{h_{\beta_n}(\cdot)\}$, with $\beta_n \uparrow 1$, is uniformly bounded and equicontinuous; thus a similar approach as in [148] can be followed. Georgin [72], [73] also explored this approach, under some ergodicity conditions. Tijms [179] and Hübner [96] directly studied the ACOE, under some ergodicity assumptions, by showing that the undiscounted dynamic programming map is a contraction on $\mathcal{M}_b(\mathcal{S})$, with respect to the span seminorm. For an excellent presentation of these methods and the type of ergodicity conditions used, see [82, §3.3]. Wijngaard [201], [202] and Kumar [101] studied the problem under Doeblin's condition using an operator theoretic method. Under several conditions, Kurano [104] obtained solutions to the ACOE and also showed the existence of an average optimal stationary deterministic policy. Also, in [105]–[107], he obtained the existence of an optimal stationary deterministic policy under Doeblin's condition. For a comprehensive presentation of the different recurrence conditions used for the above purposes, see [86].

The study of *partially observable controlled Markov processes* (POCMP) was initiated independently by various authors [5], [46], [50], [161], [162]. The reduction to models with complete information (see §7) was exhibited for various cases in [5], [138], [151], [205]. The study of finite state space POCMP with an AC criterion was initiated by Sondik [170]. Transforming the problem into an equivalent *completely observable* (CO) problem with Borel state space, Sondik tried to cast the problem in the framework of Ross [148] but did not show equicontinuity of $\{h_\beta(\cdot)\}_{\beta \in (0,1)}$. Ross [150], Wang [189], and White [191] showed this equicontinuity condition for specific scalar replacement problems. Ohnishi, Mine, and Kawai [132] studied a multistate replacement problem by using concavity properties of $h_\beta(\cdot)$. Platzman studied the general problem of finite state and action space POCMP, also by using concavity properties of the functions $h_\beta(\cdot)$. Under certain reachability conditions, he proved that the family $\{h_\beta(\cdot)\}_{\beta \in (0,1)}$ is uniformly bounded. However, even though this family may not be equicontinuous with respect to the Euclidean metric, he showed that it is equi-Lipschitzian with respect to some other appropriate metric, thus putting the problem within the framework of Ross [148]. Fernández-Gauchera, Arapostathis, and Marcus [62], [63] followed a different approach to the problem, using the concepts of invariant sets of a CMP and controlled sub-Markov processes. This approach allows us to consider POCMP with countable state and observation spaces. Borkar [26] also studied the problem via his convex analytic approach.

4. Finite state space. In this section, we will consider models with a finite state space. Initially, we restrict our attention to the case when \mathcal{A} is a finite set; models with compact action space will be discussed at the end of the section.

4.1. Finite action spaces. Let $\mathcal{S} = \{1, \dots, k\}$. In this case, Π_{SD} is finite. This fact plays a crucial role in the analysis for the average cost problem. For a policy $\pi \in \Pi$, let $J_\beta(\pi)$ denote the vector $(J_\beta(1, \pi), \dots, J_\beta(k, \pi))^T$; similarly, we define $J_N(\pi)$, $J(\pi)$,

J_β^* , J^* , and J_N^* . For a stationary deterministic policy $f \in \Pi_{SD}$, let $P(f)$ denote the transition matrix of the corresponding process and

$$c(f) := (c(1, f(1)), \dots, c(k, f(k)))^T.$$

Also, the (i, j) th entry in the n th power of the transition matrix $P(f)$ will be denoted by $P_{ij}^n(f)$ or $P^n(f)(i, j)$. It is well known that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n(f) := P^*(f)$$

exists [18], [87, Chap. 4], [137], where $P^0(f) = I$ (the $k \times k$ identity matrix). Using the theory of stochastic matrices, the following results can be proved. For details, see [8], [18], [87], [137].

THEOREM 4.1. *For each $f \in \Pi_{SD}$,*

- (i) $J(f) = P^*(f)c(f)$;
- (ii) *The number of linearly independent equations in $(I - P(f))w = c(f) - J(f)$ is k minus the number of communicating classes in $P(f)$;*
- (iii) *The equations*

$$(4.1) \quad (I - P(f))w = c(f) - v,$$

$$(4.2) \quad P^*(f)w = 0$$

have solutions $v = J(f)$ and $w = w(f)$, where

$$w(f) := (I - P(f) + P^*(f))^{-1}(I - P^*(f))c(f);$$

- (iv) *$v = J(f)$ and $w = w(f)$ are the unique solutions to (4.1) and (4.2) for which $v(s) = v(s')$ if s and s' are in the same communicating class of $P(f)$, and $v(s) = J(s, f)$ if state s is transient in $P(f)$.*

Remark 4.1. (a) It is easily seen from the above theorem that if, under an $f \in \Pi_{SD}$, the process is irreducible or unichain (see [87]), then $J(\cdot, f)$ is constant.

- (b) The matrix

$$H(f) := (I - P(f) + P^*(f))^{-1}(I - P^*(f))$$

is called the *deviation matrix*. It plays a fundamental role in the analysis. For the discounted case, $J_\beta(f) = (I - \beta P(f))^{-1}c(f)$. Analogous results can be developed for the average cost case using $H(f)$. The following result, due to Miller and Veinott [127] and Lamond and Puterman [110], can be proved using the spectral theory of stochastic matrices.

THEOREM 4.2. *Let $\beta \in [0, 1)$ and $\lambda = (1 - \beta)\beta^{-1}$. Let $f \in \Pi_{SD}$ and ν be the eigenvalue of $P(f)$ less than one with largest modulus. If $0 \leq \lambda \leq 1 - |\nu|$, then*

$$(4.3) \quad (\lambda I + I - P)^{-1} = \lambda^{-1}P^*(f) + \sum_{n=0}^{\infty} (-\lambda)^n H^{n+1}(f)$$

and

$$(4.4) \quad J_\beta(f) = (1 + \lambda) \left[\lambda^{-1}P^*(f)c(f) + \sum_{n=0}^{\infty} (-\lambda)^n H^{n+1}(f)c(f) \right].$$

Remark 4.2. (a) The quantity $h(f) := H(f)c(f)$ plays a crucial role in the analysis of the problem. It is called the *bias* or *transient cost*. It can be easily seen from the Neumann series expansion of $(I - P(f) + P^*(f))^{-1}$ [137] that, for $s \in \mathcal{S}$,

$$h(f)(s) = E_s^f \left[\sum_{t=0}^{\infty} (c(X_t, f(X_t)) - J(X_t, f)) \right].$$

From the above representation, $h(f)$ can be interpreted as the expected total cost for a CMP with cost $c - J$. If $P(f)$ is aperiodic, the distribution of X_t converges to a limiting distribution, so eventually $c(X_t, f(X_t))$ and $J(X_t, f)$ will differ very little. Thus, $h(f)$ can be thought of as the expected total cost “until convergence” or the expected total cost during the “transient” phase of the evolution of the process [137].

(b) Howard [95] has shown that

$$J_N(f) = NJ(f) + h(f) + o(1).$$

Therefore, as N becomes large, for each $s \in \mathcal{S}$, $J_N(f)$ approaches a straight line with slope $J(f)$ and intercept $h(f)$. When $J(f)(s)$ is constant, $J_N(s) - J_N(s')$ approaches $h(f)(s) - h(f)(s')$, so that $h(f)$ is the asymptotic relative difference of starting the process in two states s and s' . That is why $h(f)$ is often referred to as the relative value. See [14, pp. 304–308], [36] for a good discussion of these matters.

(c) Expansion (4.4) extends Blackwell’s expansion [18].

(d) Using expansion (4.4), the following important result is immediate.

COROLLARY 4.1. For $f \in \Pi_{SD}$, $J(f) = \lim_{\beta \uparrow 1} (1 - \beta)J_\beta(f)$.

Following Blackwell [18] and Derman [40], we now prove the following existence results.

THEOREM 4.3. *There exists an $f \in \Pi_{SD}$ that is discount optimal for all β sufficiently close to 1 and is also optimal for the average cost criterion.*

Proof. For each $f \in \Pi_{SD}$ and $s \in \mathcal{S}$, $J_\beta(s, f)$ is obviously an analytic function of β . Let $\{\beta_n\}$, $0 < \beta_n < 1$ be a sequence such that $\beta_n \uparrow 1$. For a fixed n , let $f_n \in \Pi_{SD}$ be β_n -discount optimal (see Theorem 2.1). Since Π_{SD} is a finite set, the sequence $\{f_n\}$ must contain at least one $f^* \in \Pi_{SD}$ that occurs infinitely often. Let $\{\beta_{n_k}\}$ be a subsequence of $\{\beta_n\}$ such that $\beta_{n_k} \uparrow 1$ and $f^* = f_{n_1} = f_{n_2} = \dots$. Then, for every $g \in \Pi$, $J_{\beta_{n_k}}(f^*) \leq J_{\beta_{n_k}}(g)$. Since all coordinates of $J_\beta(f^*)$ and $J_\beta(g)$ are analytic functions of β , it follows that

$$J_\beta(f^*) \leq J_\beta(g)$$

for all β near 1. Since this holds for all $g \in \Pi$, it follows that f^* is β -discount optimal for all β near 1. We next show that f^* is average optimal. Let $\pi \in \Pi$. Then

$$(1 - \beta_{n_k})J_{\beta_{n_k}}(f^*) \leq (1 - \beta_{n_k})J_{\beta_{n_k}}(\pi), \quad k = 1, 2, \dots$$

Therefore, letting $k \rightarrow \infty$ and using Theorem 4.1 and a standard Tauberian theorem (Theorem A.2 in the Appendix), it follows that

$$\begin{aligned} J(f^*) &= \lim_{\beta \uparrow 1} (1 - \beta)J_\beta(f^*) \\ &= \lim_{k \rightarrow \infty} (1 - \beta_{n_k})J_{\beta_{n_k}}(f^*) \\ &\leq \limsup_{k \rightarrow \infty} (1 - \beta_{n_k})J_{\beta_{n_k}}(\pi) \leq J(\pi), \end{aligned}$$

and the proof is complete. \square

We now briefly mention three numerical approaches. For details, we refer to [14], [137], [180], among others. Our presentation follows [137].

Value iteration. We assume that, under any $f \in \Pi_{SR}$, the corresponding chain is unichain and aperiodic. For any positive integer N , the finite horizon value function J_N^* satisfies the equation

$$(4.5) \quad J_{N+1}^* = \min_{f \in \Pi_{SD}} \{c(f) + P(f)J_N^*\}.$$

Equation (4.5) can act as an iteration equation with $J_0^* \equiv 0$ as the initial condition. Let $f_{N+1}^* \in \Pi_{SD}$ realize the minimum in (4.5). We can treat $(1/N)J_N^*$ and f_N^* as our guesses for J^* and an average optimal policy. Then $J_N^* - NJ^*$ converges as $N \rightarrow \infty$. Also, there exists an integer N_0 such that, for any $N \geq N_0$, any $f \in \Pi_{SD}$ that attains the minimum in (4.5) is average optimal. However, this property does not yield an error estimate and hence fails to provide a stopping rule for the iteration scheme. To this end, with $h = (h(1), \dots, h(k))$, we let

$$L(h) := \min_{x \in \mathbf{S}} \{Th(x) - h(x)\}, \quad U(h) := \max_{x \in \mathbf{S}} \{Th(x) - h(x)\}.$$

It can be shown that [137]

$$\min_{x \in \mathbf{S}} \{J_N^*(x) - J_{N-1}^*(x)\} \leq J^* \leq \max_{x \in \mathbf{S}} \{J_N^*(x) - J_{N-1}^*(x)\}$$

and

$$L(J_{N-1}^*) \leq L(J_N^*) \leq J^* \leq U(J_N^*) \leq U(J_{N-1}^*).$$

Furthermore, $\lim_{N \rightarrow \infty} \{U(J_N^*) - L(J_N^*)\} = 0$. Thus, an average ε -optimal policy can be found by stopping the value iteration when

$$U(J_N^*) - L(J_N^*) < \varepsilon.$$

There are other variants of this approach; see [54], [56], and [96].

Linear programming. To simplify our presentation, we will assume that, under any $f \in \Pi_{SR}$, the corresponding process is irreducible. Let $P(f)$ denote the transition matrix of the process, and $\eta(f) \in \mathcal{P}(\mathbf{S})$ its invariant measure. Then, for any $s \in \mathbf{S}$, $J(s, f) = J(f)$, a constant, and

$$J(f) = \sum_{s \in \mathbf{S}} \sum_{a \in U(s)} c(s, a) f(s, a) \eta(f)(s).$$

Therefore, the average cost problem can be reduced to the following linear programming problem:

$$(4.6a) \quad \text{minimize} \quad \sum_{s \in \mathbf{S}} \sum_{a \in U(s)} c(s, a) x(s, a)$$

subject to

$$(4.6b) \quad x(s, a) \geq 0, \quad s \in \mathbf{S}, \quad a \in U(s),$$

$$(4.6c) \quad \sum_{s \in \mathbf{S}} \sum_{a \in U(s)} x(s, a) = 1,$$

$$(4.6d) \quad \sum_{a \in U(s)} x(s, a) = \sum_{s' \in \mathbf{S}} \sum_{a \in U(s')} x(s', a) P(s' | s, a).$$

Under the irreducibility assumption, the simplex method can be employed to find an optimal stationary deterministic policy. This formulation is due to Manne [125].

Policy improvement. We work under the irreducibility assumption. The dual to the linear program (4.6a)–(4.6d) is the problem of finding variables g and $h(s)$, $s \in \mathbf{S}$, to

$$(4.7a) \quad \text{maximize } g$$

subject to

$$(4.7b) \quad g + \sum_{s' \in \mathbf{S}} (\delta(s, s') - P(s' | s, a)) h(s) \leq c(s, a),$$

$(s, a) \in \mathbf{S} \times U(s)$, where $\delta(s, s')$ is the Kronecker delta.

The functional equation

$$(4.8) \quad g + h(s) = \min_{a \in U(s)} \left\{ c(s, a) + \sum_{s' \in \mathbf{S}} P(s' | s, a) h(s') \right\}$$

is equivalent to (4.7a), (4.7b) under the irreducibility assumption and is the average cost optimality equation [87]. We will discuss this equation in detail in the next section. It will be shown that an $f \in \Pi_{SD}$ is optimal if and only if f realizes the pointwise minimum in (4.8), and then g is the optimal average cost. This suggests the following iteration algorithm.

- (i) Let $n = 1$. Choose $f_n \in \Pi_{SD}$. Let $h_n(s) \equiv 0$ for all $s \in \mathbf{S}$.
- (ii) Find a solution g_n and $h_n(s)$ of the following equation:

$$g_n + h_n(s) = c(s, f_n(s)) + \sum_{s' \in \mathbf{S}} P(s' | s, f_n(s)) h_n(s').$$

- (iii) For each $s \in \mathbf{S}$, compute

$$\phi_n(s) = \min_{a \in U(s) \setminus \{f_n(s)\}} \left\{ c(s, a) + \sum_{s' \in \mathbf{S}} P(s' | s, a) h_n(s') \right\} - g_n - h_n(s).$$

If $\phi_n(s) \geq 0$ for all $s \in \mathbf{S}$, then f_n is average optimal and g_n is the optimal average cost. If $\phi_n(s) < 0$ for some $s \in \mathbf{S}$, then pick $a \in U(s)$ such that

$$c(s, a) + \sum_{s' \in \mathbf{S}} P(s' | s, a) h_n(s') - g_n - h_n(s) < 0.$$

Define $f_{n+1} \in \Pi_{SD}$ as $f_{n+1}(s) = a$ and $f_{n+1}(\cdot) = f_n(\cdot)$, otherwise. Then f_{n+1} yields a lower average cost. Since Π_{SD} is finite, the policy improvement scheme converges in a finite number of steps.

4.2. Compact action spaces. We now consider the problem where the action set \mathbf{A} is not finite but a compact metric space. In this situation, an optimal policy may not exist; see [51, p. 178, Ex. 1]. Note that here Π_{SD} is no longer finite. Under certain ergodicity assumptions, Martin-Löf [126] and Feinberg [57] have proved the existence of an optimal $f \in \Pi_{SD}$. We will discuss various ergodicity assumptions on a countable state space in detail in the next section. First, we focus on ε -optimal policies established by Chitashvili [35] and Feinberg [58]; see [51, Chap. 7].

THEOREM 4.4. *Under Assumptions 2.1–2.3, for every $\varepsilon > 0$, there exists an ε -optimal $f \in \Pi_{SD}$.*

Proof (Sketch). For $f \in \Pi_{SD}$, let $J(f)$ be as in Theorem 4.1. For $i \in \mathbf{S}$, let

$$(4.9) \quad \tilde{J}(i) = \inf_{f \in \Pi_{SD}} J(f)(i).$$

Clearly, $J^*(i) \leq \tilde{J}(i)$, for each $i \in \mathbf{S}$. Corresponding to $i \in \mathbf{S}$, select an $f_i \in \Pi_{SD}$ such that

$$(4.10) \quad J(f_i)(i) \leq \tilde{J}(i) + \varepsilon.$$

The set $\tilde{\mathbf{A}} = \{f_i(j) : i, j \in \mathbf{S}\}$ is obviously finite. Taking the action set to be $\tilde{\mathbf{A}}$, the preceding results can be applied to the finite CMP $(\mathbf{S}, \tilde{\mathbf{A}}, P, c)$. For this model, there exists a stationary deterministic policy, say f^* , which is average optimal. Thus

$$(4.11) \quad J(f^*)(i) \leq J(f_i)(i) \leq \tilde{J}(i) + \varepsilon \quad \text{for each } i \in \mathbf{S}.$$

Let

$$\rho^*(i) := \limsup_{\beta \uparrow 1} (1 - \beta)J_\beta^*(i).$$

Then, by Theorem A.2, in the Appendix,

$$(4.12) \quad \rho^*(i) \leq J^*(i) \quad \text{for each } i \in \mathbf{S}.$$

By (4.11), it suffices to show that $J^*(i) = \tilde{J}(i)$, for each $i \in \mathbf{S}$. From (4.12), it then suffices to show that $\rho^*(i) \geq \tilde{J}(i)$. For each $\beta \in (0, 1)$, let $f_\beta \in \Pi_{SD}$ be β -discount optimal. Let f be a limit point of f_β as $\beta \uparrow 1$. Then using (4.4) (which is valid in this case as well) and Assumptions 2.1–2.3, it can be shown that

$$\rho^*(i) \geq J(f)(i) \geq \tilde{J}(i). \quad \square$$

Concerning the existence of an optimal policy, we state the following result.

THEOREM 4.5. *Let Assumptions 2.1 and 2.2 hold and further assume that $c(\cdot)$ is continuous on Π_{SR} and that, under any $f \in \Pi_{SR}$, the corresponding chain is unichain. Then there exists an optimal policy in $f \in \Pi_{SD}$.*

The result is almost immediate from the fact that, under the unichain assumption, $P^*(\cdot)$, and therefore also $J(\cdot)$, is continuous on Π_{SR} [91, Lemma 10.2]. For further details, including the convergence of a policy improvement algorithm, see [94].

5. Countable state space. The average cost problem becomes much more complicated when the state space is countable. Maitra [121] has given a counterexample that shows that there need not exist an optimal policy. In [122] Maitra has studied a particular problem in which there does not exist any policy that is β -discount optimal for all β sufficiently close to 1. Flynn [68] has constructed a more dramatic counterexample. In his example, there exists an average optimal policy in Π_{SD} . Nevertheless he exhibits an $f \in \Pi_{SD}$ and a $\beta_0 \in (0, 1)$ such that f is β -discount optimal for all $\beta \in (\beta_0, 1)$, but it is not average optimal. Fisher and Ross [67] have presented a counterexample that shows that the optimal policy need not be stationary or deterministic. We refer to [150] for several other counterexamples. It is apparent that the average

cost problem is closely related to the ergodic behavior of the process, and it is well known that the ergodic theory of Markov processes on a countable state space is much more involved than on a finite state space; for example, a Markov process on a finite state space cannot be null recurrent. Another vital difference in this case is that the number of stationary deterministic policies is no longer finite. To study the ergodic theory, some recurrence conditions are necessary. There are many such conditions available in the literature [26], [52], [178]; we will survey a few representative ones.

In what follows, the state space $\mathbf{S} = \{0, 1, 2, \dots\}$. For each $i \in \mathbf{S}$, the action space $U(i)$ is a prescribed compact metric space. We will always assume that, for fixed $i, j \in \mathbf{S}$, $c(i, \cdot)$, $P(i | j, \cdot)$, are continuous. These conditions can be weakened or dropped in several places, as will be clear from the specific context.

Derman [38] studied the ACOE that, with ρ a scalar and $h : \mathbf{S} \rightarrow \mathbb{R}$, takes the following form:

$$(5.1) \quad \rho + h(i) = \min_{a \in U(i)} \left\{ c(i, a) + \sum_{j \in \mathbf{S}} P(j | i, a) h(j) \right\}.$$

A solution to (5.1) is a pair (ρ, h) satisfying it.

Suppose that $f \in \Pi_{SD}$ is a minimizing selector in (5.1). Then (5.1) becomes

$$(5.1') \quad \rho + h(i) = c(i, f(i)) + \sum_{j \in \mathbf{S}} P(j | i, f(i)) h(j).$$

Equation (5.1') asserts that, apart from ρ , the cost if the process stops now equals the expected cost if it continues under the policy f for just one more period. We can give a similar interpretation to (5.1). Hence, we may think that ρ is the average cost under f and that no other $f \in \Pi_{SD}$ has a smaller average cost. Thus, the function h in (5.1) is roughly a measure of how much we are prepared to pay to stop the process, though continuing to pay an average cost ρ in the future [141] (cf. Remark 4.2(a)). Therefore, the function h may be viewed as a cost potential. Also, by a stochastic representation of h , using (5.1) and (5.1'), h is indeed a potential. Hordijk [91] has pursued this line of thought in great detail, which we will discuss later.

We start with a characterization of optimal policies.

THEOREM 5.1. *If the ACOE has a solution (ρ, h) satisfying*

$$(5.2) \quad \lim_{t \rightarrow \infty} \frac{1}{t} E_i^\pi h(X_t) = 0 \quad \forall \pi \in \Pi_{SD}, \quad \forall i \in \mathbf{S},$$

then there exists an $f \in \Pi_{SD}$ such that

$$\rho = J(i, f) = J^*(i) \quad \forall i \in \mathbf{S}.$$

Moreover, an $f \in \Pi_{SD}$ is average optimal if, for each $i \in \mathbf{S}$,

$$(5.3) \quad c(i, f(i)) + \sum_{j \in \mathbf{S}} P(j | i, f(i)) h(j) = \min_{a \in U(i)} \left\{ c(i, a) + \sum_{j \in \mathbf{S}} P(j | i, a) h(j) \right\},$$

and, conversely, if an $f \in \Pi_{SD}$ is average optimal and the corresponding chain is irreducible and positive recurrent, then (5.3) holds.

Proof. Let $f \in \Pi_{SD}$ satisfy (5.3). Then, since

$$E_i^f [h(X_{t+1}) | \mathfrak{F}_t] = \sum_{j \in \mathbf{S}} P(j | X_t, f(X_t)) h(j),$$

it follows from (5.1) and (5.3) that

$$(5.4) \quad \rho + h(X_t) = c(X_t, f(X_t)) + E_i^f[h(X_{t+1}) \mid \mathfrak{F}_t].$$

Summing (5.4) from $t = 0$ to $N - 1$, dividing by N , and taking expectations, we obtain

$$\rho = \frac{1}{N} E_i^f \left[\sum_{t=0}^{N-1} c(X_t, f(X_t)) \right] + \frac{E_i^f[h(X_N)] - h(i)}{N}.$$

Next, letting $N \rightarrow \infty$ and using (5.2) yields

$$\rho = \lim_{N \rightarrow \infty} \frac{1}{N} E_i^f \left[\sum_{t=0}^{N-1} c(X_t, f(X_t)) \right].$$

On the other hand, if π is any other policy, we can show using the same arguments that

$$\rho \leq \limsup_{N \rightarrow \infty} \frac{1}{N} E_i^\pi \left[\sum_{t=0}^{N-1} c(X_t, A_t) \right].$$

Hence, f is average optimal. Conversely, let $f \in \Pi_{SD}$ be average optimal and suppose that the corresponding chain is irreducible and positive recurrent. If f does not satisfy (5.3), then there exist $i_0 \in \mathcal{S}$, $a_0 \in U(i_0)$ and $\delta > 0$ such that

$$(5.5) \quad \begin{aligned} c(i_0, f(i_0)) + \sum_{j \in \mathcal{S}} P(j \mid i_0, f(i_0)) h(j) \\ = c(i_0, a_0) + \sum_{j \in \mathcal{S}} P(j \mid i_0, a_0) h(j) + \delta. \end{aligned}$$

Let $f' \in \Pi_{SD}$ be defined as follows:

$$f'(i) = \begin{cases} f(i) & \text{if } i \neq i_0, \\ a_0 & \text{if } i = i_0. \end{cases}$$

Then, using (5.5) along with irreducibility and positive recurrence, it is easily seen that $J(i_0, f') < J(i_0, f)$, which contradicts the average optimality of f . \square

Remark 5.1. (a) We say that (5.1) admits a bounded solution if $h(\cdot)$ is bounded. If the ACOE has a bounded solution, then (5.2) is clearly satisfied; moreover, using the martingale stability theorem [117, p. 53], it can be shown that the $f \in \Pi_{SD}$ selecting the minimum in (5.3) is sample path average optimal [72].

(b) Various extensions of last assertion of Theorem 5.1 have been obtained by Sennott [158].

Derman and Veinott [43] have prescribed a certain recurrence condition that ensures that (5.1) admits a bounded solution. We will discuss it later in this section. The ACOE resembles the dynamic programming equation, and Theorem 5.1 is analogous to a dynamic programming characterization of an optimal policy. However, the dynamic programming heuristic does not lead directly to the ACOE. Taylor [177] developed a vanishing discount approach for a particular problem, which was extended for the general case by Ross [147]–[150]. Our presentation here follows Ross [150]. As noted earlier, the average case can in some sense be treated as the limiting case of

the discounted problem as the discount factor approaches 1. The discounted value function $J_\beta^*(\cdot)$ satisfies the DCOE (cf. Theorem 2.1)

$$J_\beta^*(i) = \min_{a \in U(i)} \left\{ c(i, a) + \beta \sum_{j \in \mathcal{S}} P(j | i, a) J_\beta^*(j) \right\},$$

and a β -discounted optimal policy selects a minimizing action. One possible way of finding an average optimal policy might be to choose the actions minimizing

$$\lim_{\beta \rightarrow 1} \left\{ c(i, a) + \beta \sum_{j \in \mathcal{S}} P(j | i, a) J_\beta^*(j) \right\}.$$

However, this limit need not exist and indeed would often be infinite for all actions. The situation can nevertheless be salvaged by considering a “differential” discounted value function, i.e., $h_\beta(i) := J_\beta^*(i) - J_\beta^*(0)$, where $0 \in \mathcal{S}$ is an arbitrary, fixed state. The function $h_\beta(\cdot)$ satisfies

$$(5.6) \quad (1 - \beta)J_\beta^*(0) + h_\beta(i) = \min_{a \in U(i)} \left\{ c(i, a) + \beta \sum_{j \in \mathcal{S}} P(j | i, a) h_\beta(j) \right\}.$$

From (5.6) it is now apparent that (5.1) can be derived under certain conditions by letting $\beta \rightarrow 1$. We state here a simple result [150], despite the fact that it also holds under weaker hypotheses (see Theorem 5.9).

THEOREM 5.2. *Suppose that there exists a constant $K > 0$ such that $|h_\beta(i)| \leq K$, for all $\beta \in (0, 1)$ and $i \in \mathcal{S}$. Then*

- (i) *The ACOE admits a bounded solution (ρ, h) ;*
- (ii) *For some sequence $\beta_n \rightarrow 1$, $h(i) = \lim_{n \rightarrow \infty} h_{\beta_n}(i)$, $i \in \mathcal{S}$;*
- (iii) *$\lim_{\beta \rightarrow 1} (1 - \beta)J_\beta^*(i) = \rho$ for any $i \in \mathcal{S}$.*

Proof. Let $\beta_n \uparrow 1$ be given. By the uniform boundedness of $h_\beta(\cdot)$, using a diagonalization procedure, we can find a subsequence, which for simplicity we also denote by β_n , such that $h_{\beta_n}(i) \rightarrow h(i)$ for each $i \in \mathcal{S}$, where $h(\cdot)$ is a bounded function. Again, since $(1 - \beta_n)J_{\beta_n}^*(0)$ is bounded, there is a further subsequence $\beta_{n_k} \uparrow 1$ such that

$$\lim_{k \rightarrow \infty} (1 - \beta_{n_k})J_{\beta_{n_k}}^*(0)$$

exists. Part (i) of the theorem then follows from (5.6) and an application of the dominated convergence theorem. Furthermore, by Theorem 5.1, ρ is the minimum average cost. Since the above results are independent of the sequence chosen, (iii) then follows. \square

Remark 5.2. It has been shown [64] that, if the ACOE has a bounded solution, then there exists a constant $K > 0$ such that $|h_\beta(i)| \leq K$ for all $\beta \in (0, 1)$, $i \in \mathcal{S}$.

5.1. Bounded costs. In this section, we assume that $c(\cdot, \cdot)$ is bounded. Ross [150] has proved that under a Derman–Veinott [43] type recurrence condition (see (5.7), below), the uniform boundedness hypothesis of Theorem 5.2 is satisfied.

THEOREM 5.3. *Let $f \in \Pi_{SD}$ and let $\{X_t\}$ be the corresponding state process. Let*

$$\tau = \min\{t \geq 1 : X_t = 0\}.$$

If there exists a $K > 0$ such that

$$(5.7) \quad E_i^f[\tau] < K$$

for all $f \in \Pi_{SD}$ and all $i \in \mathbf{S}$, then $h_\beta(i)$ is bounded uniformly in $\beta \in (0, 1)$ and $i \in \mathbf{S}$.

Proof. Let $\beta \in (0, 1)$ and $f_\beta \in \Pi_{SD}$ be β -discount optimal. We have

$$(5.8) \quad \begin{aligned} J_\beta^*(i) &= E_i^{f_\beta} \left[\sum_{t=0}^{\infty} \beta^t c(X_t, f_\beta(X_t)) \right] \\ &= E_i^{f_\beta} \left[\sum_{t=0}^{\tau-1} \beta^t c(X_t, f_\beta(X_t)) \right] + E_i^{f_\beta} \left[\sum_{t=\tau}^{\infty} \beta^t c(X_t, f_\beta(X_t)) \right] \\ &\leq M E_i^{f_\beta}[\tau] + J_\beta^*(0) E_i^{f_\beta}[\beta^\tau], \end{aligned}$$

where M is a bound on $c(\cdot, \cdot)$. From (5.7) and (5.8), it follows that

$$(5.9) \quad J_\beta^*(i) - \beta J_\beta^*(0) \leq MK.$$

Also, from (5.8) and applying Jensen's inequality, we obtain

$$J_\beta^*(i) \geq J_\beta^*(0) E_i^{f_\beta}[\beta^\tau] \geq J_\beta^*(0) \beta^K.$$

Therefore,

$$(5.10) \quad \begin{aligned} J_\beta^*(0) - J_\beta^*(i) &\leq (1 - \beta^K) J_\beta^*(0) \\ &\leq (1 - \beta^K) \frac{M}{1 - \beta} \leq MK. \end{aligned}$$

The desired result follows from (5.9) and (5.10). \square

After the work of Derman [38], Derman and Veinott [43], and Ross [147], [148], several recurrence conditions have appeared [178]. We explore a few representative ones.

Let $f \in \Pi_{SD}$. For a finite set $A \subset \mathbf{S}$, let

$$(5.11) \quad \tau_A = \min\{t \geq 1 : X_t \in A\}.$$

Assumption 5.1. There is a finite $A \subset \mathbf{S}$ and a constant $K > 0$ such that $E_i^f[\tau_A] < K$ for all $i \in \mathbf{S}$ and $f \in \Pi_{SD}$. Furthermore, for any $f \in \Pi_{SD}$ the corresponding process does not have two disjoint invariant sets.

Assumption 5.2. There exists a constant $K > 0$, and, for every $f \in \Pi_{SD}$, there is a state $j(f) \in \mathbf{S}$ such that

$$E_i^f[\tau_{\{j(f)\}}] < K \quad \forall i \in \mathbf{S}.$$

Assumption 5.3 (simultaneous Doeblin). There is a finite set A , an integer $n \geq 1$ and a scalar $\alpha > 0$ such that

$$\sum_{j \in A} P(j \mid i, f(i)) \geq \alpha$$

for all $i \in \mathbf{S}$ and all $f \in \Pi_{SD}$. Furthermore, for any $f \in \Pi_{SD}$, the corresponding process does not have two disjoint invariant sets.

Assumption 5.4 (scrambling). There is an integer $n \geq 1$ and a scalar $\alpha > 0$ such that, for any $f \in \Pi_{SD}$,

$$\sum_{j \in \mathbf{S}} \min\{P_{i_1,j}^n(f), P_{i_2,j}^n(f)\} \geq \alpha \quad \forall i_1, i_2 \in \mathbf{S}.$$

Assumption 5.5 (ergodicity). There is an integer $n \geq 1$ and a scalar $\rho > 0$ such that, for each $f \in \Pi_{SD}$, there exists an $\eta(f) \in \mathcal{P}(\mathbf{S})$ for which

$$\sum_j |P_{ij}^m(f) - \eta(f)(j)| \leq 2(1 - \rho)^{\lfloor m/n \rfloor}$$

for all $i \in \mathbf{S}$ and $m \geq 1$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x .

Remark 5.3. Clearly Assumptions 5.1 and 5.2 are generalizations of the Derman–Veinott condition. Hordijk [91] has proved the existence of a bounded solution to the ACOE using Assumption 5.1. Under Assumption 5.5, for each $f \in \Pi_{SD}$, $\eta(f)$ is the unique invariant measure of the corresponding process.

Federgruen, Hordijk, Tijms [52] have established the following theorem.

THEOREM 5.4. *Assumptions 5.1–5.3 are equivalent. Also, if for any $f \in \Pi_{SD}$ the corresponding process is aperiodic, then Assumptions 5.1–5.5 are equivalent.*

Remark 5.4. Under any one of Assumptions 5.1–5.5, Federgruen, Hordijk, and Tijms [52] have established the existence of a bounded solution to the ACOE by extending the vanishing discount approach of Taylor and Ross.

We have thus far seen several recurrence conditions which are sufficient for the ACOE to admit a bounded solution. Cavazos-Cadena [30], [31] has dealt with the converse question of what are the necessary recurrence conditions for the ACOE to have a bounded solution. He has obtained the following result. Consider the following assumption.

Assumption 5.6. There exists a constant $K > 0$ such that, for each bounded and measurable $c : \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$ and every collection $\{U(i) : i \in \mathbf{S}\}$, $U(i) \subset \mathbf{A}$, there exist $\rho \in \mathbb{R}$ and $h : \mathbf{S} \rightarrow \mathbb{R}$ bounded that solve (5.1) and satisfy $\|h\| \leq K\|c\|$, where $\|\cdot\|$ is the sup norm.

THEOREM 5.5. *Assumptions 5.2 and 5.6 are equivalent.*

The proof follows by an application of the uniform boundedness principle. For details and other variants, we refer to [30], [31]. Thus, an assumption on the existence of a bounded solution to the ACOE necessarily imposes a strong recurrence structure on the system. Also, note that Assumption 5.6 involves not just one CMP but a family of CMP (one for each c and $\{U(i)\}$). Since it is equivalent to Assumptions 5.1–5.3 and under aperiodicity conditions to Assumptions 5.1–5.5, it follows that Assumptions 5.1–5.5 are too strong for many important applications. In fact, there are interesting situations [20] in which these conditions are not satisfied, but for which we can find average optimal stationary deterministic policies.

Ross [148] has proved that, under the following recurrence condition, the AC can be reduced to an appropriate DC. Therefore, in view of Theorem 2.1, the problem can be resolved in this case.

THEOREM 5.6. *If there exists a constant $\alpha > 0$ such that*

$$P(0 \mid i, a) \geq \alpha > 0$$

for all $i \in \mathbf{S}$, $a \in U(i)$, then the AC can be reduced to an appropriate DC.

Proof. Let

$$\tilde{P}(j | i, \cdot) = \begin{cases} (1 - \alpha)^{-1} P(j | i, \cdot) & \text{for } j \neq 0, \\ (1 - \alpha)^{-1} (P(0 | i, \cdot) - \alpha) & \text{for } j = 0. \end{cases}$$

Let $\tilde{J}_\beta^*(\cdot)$ denote the β -discounted value function for the CMP with cost $c(\cdot, \cdot)$ and transition law $\tilde{P}(\cdot | \cdot, \cdot)$. Then it is easily verified that, for each $i \in \mathbf{S}$,

$$\alpha \tilde{J}_{1-\alpha}^*(0) + \tilde{J}_{1-\alpha}^*(i) = \min_{a \in U(i)} \left\{ c(i, a) + \sum_{j \in \mathbf{S}} P(j | i, a) \tilde{J}_{1-\alpha}^*(j) \right\}.$$

Let $f \in \Pi_{SD}$ be $(1 - \alpha)$ -discount optimal for the modified CMP. It follows from Theorem 5.1 that f is AC-optimal for the original CMP, and the optimal average cost is $\alpha \tilde{J}_{1-\alpha}^*(0)$. \square

Remark 5.5. Note that, if the ACOE has a bounded solution (ρ, h) , then ρ is the optimal average cost for any initial condition. Hence, the existence of a bounded solution to the ACOE suggests that some kind of “unchainedness” is in effect, since, for the multichain case, the average cost would, in general, depend on the initial condition. The multichain version of the ACOE is

$$(5.12a) \quad \min_{a \in U(i)} \sum_{j \in \mathbf{S}} P(j | i, a) \rho(j) = \rho(i),$$

$$(5.12b) \quad \rho(i) + h(i) = \min_{a \in U_1(i)} \left\{ c(i, a) + \sum_{j \in \mathbf{S}} P(j | i, a) h(j) \right\},$$

where

$$(5.12c) \quad U_1(i) = \left\{ a \in U(i) : \min_{a \in U(i)} \sum_{j \in \mathbf{S}} P(j | i, a) \rho(j) = \rho(i) \right\}.$$

This equation has been studied by Zijm [208] for countable state space. For more general state spaces, it was extensively studied much earlier by Yushkevich [204] (see also [51]); this work will be discussed in the next section.

If (5.12) has a bounded solution $\rho(i)$, $h(i)$, where both ρ and h are bounded functions, then we can show, as before, that $\rho(i)$ is the optimal average cost starting from state $i \in \mathbf{S}$ and a minimizing selector in (5.12) yields an average optimal stationary deterministic policy. Under a certain “geometric convergence condition,” Zijm [208] has established the existence of a bounded solution to (5.12). Under the additional assumptions that under any stationary deterministic policy the corresponding process has at most a finite number of ergodic classes, he has shown that the geometric convergence condition is equivalent to a number of recurrence conditions of the type Assumptions 5.1–5.5.

Hordijk [91] establishes the existence of an average optimal $f \in \Pi_{SD}$ without utilizing the ACOE. Let Π_{SD} be endowed with the product topology. Then Π_{SD} is compact and metrizable. Let us consider the following assumptions.

Assumption 5.7. For each $f \in \Pi_{SD}$ and $i \in \mathcal{S}$, there exists a measure $\eta_i(f) \in \mathcal{P}(\mathcal{S})$ such that $\eta_i(f)(j) = \lim_{N \rightarrow \infty} (1/N) \sum_{n=0}^{N-1} P^n(f)(i, j)$.

Assumption 5.8. $f \mapsto \eta_i(f)$ is continuous for any $i \in \mathcal{S}$.

Assumption 5.9. For each $i \in \mathcal{S}$, $\{\eta_i(f) : f \in \Pi_{SD}\}$ is tight (for a definition of tightness, see [134, Def. 3.1, p. 28]).

Assumption 5.10. For each $f \in \Pi_{SD}$, the corresponding process is recurrent.

Assumption 5.11. For each $f \in \Pi_{SD}$, the corresponding process does not have disjoint-invariant sets.

Assumption 5.12. $\{P(f)(i, \cdot) : i \in \mathcal{S}, f \in \Pi_{SD}\}$ is tight.

It is easy to see that Assumptions 5.7 and 5.8 imply that, for each $i \in \mathcal{S}$, $\{\eta_i(f) : f \in \Pi_{SD}\}$ is compact. Hence, in particular, Assumptions 5.7 and 5.8 imply Assumption 5.9. By definition, Assumption 5.9 implies Assumption 5.7. Also, it can easily be shown that Assumptions 5.9 and 5.11 imply Assumption 5.8, and that Assumption 5.12 implies 5.9. However, Assumption 5.12 may be easier to verify.

THEOREM 5.7. *Each of the following five combinations of assumptions is sufficient for the existence of an average optimal $f \in \Pi_{SD}$: (Assumption 5.7, Assumption 5.8), (Assumption 5.9, Assumption 5.10), (Assumption 5.9, Assumption 5.11), (Assumption 5.10, Assumption 5.12), (Assumption 5.11, Assumption 5.12).*

Remark 5.6. The main idea behind the proof of this theorem can be traced back to the proof of Theorem 4.3. We give the main points and skip the details. Let $\beta_n \in (0, 1)$ be a sequence such that $\beta_n \uparrow 1$, let $f_{\beta_n} \in \Pi_{SD}$ be β_n -discount optimal, and f_∞ be a limit point of $\{f_{\beta_n}\}$ in Π_{SD} . Suppose that $\rho^*(i)$ is a scalar satisfying $(1 - \beta_n)J_{\beta_n}^*(i) \rightarrow \rho^*(i)$, for each $i \in \mathcal{S}$ (along a suitable subsequence). Then, by using Tauberian and ergodic theorems, we deduce that $J^*(i) = \rho^*(i)$ and f_∞ is average optimal under (Assumption 5.7, Assumption 5.8). Under (Assumption 5.9, Assumption 5.10), f_∞ is average optimal for initial states $i \in \tilde{\mathcal{S}} := \bigcup_i \text{supp}(\eta_i(f_\infty))$, where “supp” denotes the support. Then by Assumption 5.10 there exists an \tilde{f} such that the corresponding process starting from any $i \in \mathcal{S} \setminus \tilde{\mathcal{S}}$ reaches $\tilde{\mathcal{S}}$. Set

$$\tilde{f}(i) = \begin{cases} \bar{f}(i) & \text{if } i \notin \tilde{\mathcal{S}}, \\ f_\infty(i) & \text{if } i \in \tilde{\mathcal{S}}. \end{cases}$$

It follows that \tilde{f} is average optimal. The other cases can be dealt with in a similar manner.

5.2. Unbounded costs. We have thus far considered bounded costs only. There are practical situations (e.g., in queueing systems) where the cost is typically unbounded. We assume that $c \geq 0$ (cf. Assumption 2.1). Let us now consider the ACOE for unbounded c . Note that the boundedness of c , did not play any role in the proof of Theorem 5.1. For unbounded c , the ACOE is unlikely to admit a bounded solution.

Lippman [115], [116] has studied controlled semi-Markov processes with unbounded costs. He has placed polynomial bounds on the movement of the process in one transition. He has made a further assumption that there exists an $f \in \Pi_{SD}$ such that both the mean first passage times and mean first passage costs from any state i to state zero under the policy are finite. Moreover, if $f \in \Pi_{SD}$ is close to β -discount optimal for a sequence of discount factors, then it is AC-optimal. Lippman has employed the vanishing discount approach of Taylor and Ross to establish the existence of a solution (ρ, h) to the ACOE with h satisfying (5.2), thereby establishing the existence of an

average optimal $f \in \Pi_{SD}$. He has also given some examples from queueing systems where his conditions are satisfied. However, his condition on the β -discounted value function appears to be very difficult to verify.

Hordijk [91] has used a Lyapunov stability condition to establish the existence of an average optimal $f \in \Pi_{SD}$.

Assumption 5.13 (Lyapunov condition). Let

$$\tilde{P}(f)(i, j) = \begin{cases} P(f)(i, j), & j \neq 0, \\ 0, & j = 0. \end{cases}$$

There exists a function $w : \mathbf{S} \rightarrow \mathbb{R}_+$ such that, for all $i \in \mathbf{S}$,

- (i) $c(i, f(i)) + 1 + \sum_j \tilde{P}(f)(i, j)w(j) \leq w(i)$, for all $f \in \Pi_{SD}$;
- (ii) $\sum_j P(f)(i, j)w(j)$ is continuous in f ;
- (iii) $\lim_{n \rightarrow \infty} \sum_j \tilde{P}^n(f)(i, j)w(j) = 0$.

THEOREM 5.8. *Under the above Lyapunov condition, there exists an AC-optimal $f \in \Pi_{SD}$.*

Proof (Sketch). Let $f \in \Pi_{SD}$. For $i \in \mathbf{S}$, we define $\tau_i = \min\{t \geq 1 : X_t = i\}$, where X_t is governed by f . Then, under Assumption 5.13, using the standard techniques of stochastic Lyapunov function method [91], [108], the following results can be proved:

$$(5.13) \quad E_i^f[\tau_0] \leq w(i),$$

$$(5.14) \quad E_i^f \left[\sum_{t=0}^{\tau_0-1} c(X_t, f(X_t)) \right] \leq w(i).$$

Indeed, with $n \in \mathbb{N}$ and $n > 1$,

$$\begin{aligned} E_i^f[w(X_{n \wedge \tau_0}) \mid \mathfrak{F}_{n \wedge \tau_0}] - w(i) &= -E_i^f \left[\sum_{t=0}^{n \wedge \tau_0-1} E_i^f[w(X_{t+1}) \mid X_t] - w(X_t) \right] \\ &\leq -E_i^f[n \wedge \tau_0], \end{aligned}$$

where the last inequality is due to Assumption 5.13. Hence, $E_i^f[n \wedge \tau_0] \leq w(i)$, and, letting $n \uparrow \infty$, (5.13) follows. Also, (5.14) can be proved along the same lines. By an ergodic theorem [133],

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} E_0^f \left[\sum_{t=0}^{N-1} c(X_t, f(X_t)) \right] &= (E_0^f \tau_0)^{-1} E_0^f \left[\sum_{t=0}^{\tau_0-1} c(X_t, f(X_t)) \right] \\ &=: \rho(f). \end{aligned}$$

Let $\rho^* := \inf_{f \in \Pi_{SD}} \rho(f)$. Then $\rho^* \leq w(0)$. Define

$$h(i) = \inf_{f \in \Pi_{SD}} E_i^f \left[\sum_{t=0}^{\tau_0-1} (c(X_t, f(X_t)) - \rho^*) \right].$$

Then $h(0) = 0$. Using (5.13), (5.14), and Assumption 5.13(iii), it can be shown that (ρ^*, h) is a solution of the ACOE with h satisfying (5.2), and the desired result follows. \square

Remark 5.7. (a) Note that by Assumption 5.13(i) the cost function c does not grow faster than the Lyapunov function w . Thus, there is a restriction on the growth of c imposed by w . In CMP, $w(i) = i$, $w(i) = i^2$ are typical examples of Lyapunov functions [91]. In the latter case, for example, we can treat only those unbounded cost functions that do not grow faster than quadratic functions.

(b) Assumption 5.13(iii) is crucial in showing that the cost potential h satisfies $\lim_{t \rightarrow \infty} (1/t) E_i^f h(X_t) = 0$, for all $f \in \Pi_{SD}$, and $i \in \mathcal{S}$.

Federgruen, Hordijk, and Tijms [53] have extended Hordijk's results by replacing the single attracting point $\{0\}$ by a finite set $K \subset \mathcal{S}$. Their main assumption is the following: There exists a finite set $K \subset \mathcal{S}$ such that, for each initial state $i \in \mathcal{S}$, the suprema over the mean hitting time of K and mean hitting costs are finite. This, in turn, is equivalent to the existence of a Lyapunov function $w : \mathcal{S} \rightarrow \mathbb{R}_+$ satisfying Assumption 5.13(i), where now \tilde{P} is defined as

$$\tilde{P}(f)(i, j) = \begin{cases} P(f)(i, j), & j \notin K, f \in \Pi_{SD}, \\ 0, & j \in K. \end{cases}$$

Under the additional assumptions that Assumption 5.13(ii) and (iii) hold, and the “communication condition” that for any $f \in \Pi_{SD}$ the corresponding process has no two disjoint invariant sets, they have established the existence of a solution (ρ, h) to the ACOE by employing the vanishing discount approach and have shown that h satisfies (5.2). This work has been further extended by Federgruen, Schweitzer, and Tijms [55]. They have dropped the unchainedness assumption in [53]. Instead, they assume that any state can be reached from any other state via some policy. Under this and other conditions in [53], they have established the existence of a solution (ρ, h) to the ACOE, with h satisfying (5.2). They have deviated from the vanishing discount approach and have, instead, utilized Tychonoff's fixed point theorem in their analysis. We again note that, in all these investigations, a restrictive growth condition on the cost function is imposed, as noted in Remark 5.7.

The Lyapunov stability condition necessarily imposes a blanket stability (i.e., positive recurrence) of certain states (cf. (5.13)), which may be very restrictive. On the other hand, (5.2) is not easy to verify in general and, indeed, may not hold in the case of many queueing models [141]. Another generalization of the boundedness of the solution of the ACOE could be boundedness from below. This will be the case if the cost function has some “monotone” properties, which naturally arise in various queueing models. This line of thought has been pursued in various ways in [24], [28], [74], [76], [77], [141], [142], [155], [156], [172], [190].

Sennott [155], [156] has prescribed very general conditions in this direction. We will now briefly describe them. Consider the following assumptions.

Assumption 5.14. For every $i \in \mathcal{S}$ and every $\beta \in (0, 1)$, $J_\beta^*(i) < \infty$.

Assumption 5.15. There exists a nonnegative integer L such that

$$h_\beta(i) := J_\beta^*(i) - J_\beta^*(0) \geq -L.$$

Assumption 5.16. There exists a function $M : \mathcal{S} \rightarrow \mathbb{R}_+$ such that $h_\beta(i) \leq M(i)$ for all $i \in \mathcal{S}$ and any $\beta \in (0, 1)$. For every $i \in \mathcal{S}$, there exists an $a(i) \in U(i)$ such that

$$\sum_j P(j \mid i, a(i)) M(j) < \infty.$$

THEOREM 5.9. *Under Assumptions 5.14–5.16, there exists an AC-optimal $f \in \Pi_{SD}$.*

Proof. Let $\beta_n \in (0, 1)$ be such that $\beta_n \uparrow 1$. Let f_{β_n} be β_n -discount optimal. Let f be a limit point of f_{β_n} as $n \rightarrow \infty$. To simplify the notation, all subsequences of β_n will also be denoted by β_n . By Assumption 5.16 and a diagonal argument, there exists a function $h : \mathbf{S} \rightarrow \mathbb{R}$ such that $\lim_{n \rightarrow \infty} h_{\beta_n}(\cdot) = h(\cdot)$. By Assumption 5.15, $h(\cdot) \geq -L$. Let $\rho : \mathbf{S} \rightarrow \mathbb{R}_+$ be a function such that $\lim_{n \rightarrow \infty} (1 - \beta_n)J_{\beta_n}^*(i) = \rho(i)$. Using Assumption 5.16, it is easy to see that $\rho(i) = \rho^*$, a constant. Now, for $i \in \mathbf{S}$,

$$(5.15) \quad (1 - \beta_n)J_{\beta_n}^*(0) + h_{\beta_n}(i) = c(i, f_{\beta_n}(i)) + \beta_n \sum_{j \in \mathbf{S}} P(j \mid i, f_{\beta_n}(i)) h_{\beta_n}(j).$$

Fix an $i \in \mathbf{S}$. Add L to both sides to make $(h_{\beta_n}(i) + L) \geq 0$ and take “liminf” on both sides of (5.15). Then, by Fatou’s lemma and the assumption of continuity of $P(j \mid i, \cdot)$, we conclude that

$$\rho^* + h(i) \geq c(i, f(i)) + \sum_j P(j \mid i, f(i)) h(j).$$

Since $h(\cdot)$ is bounded below, the proof of Theorem 5.1 can be modified to show that $J(i, f) \leq \rho^*$. By Theorem A.2 in the Appendix, $J(i, \pi) \geq \rho^*$ for any $\pi \in \Pi$. Hence, $J(i, f) = J^*(i) = \rho^*$, and f is AC-optimal. \square

Remark 5.8. (a) From the above proof, it is clear that if ρ is a scalar, $h : \mathbf{S} \rightarrow \mathbb{R}$ is bounded below, and

$$(5.16) \quad \rho + h(i) \geq \min_{a \in U(i)} \left\{ c(i, a) + \sum_j P(j \mid i, a) h(j) \right\},$$

then ρ is the optimal average cost, and any $f \in \Pi_{SD}$ selecting the minimum on the right-hand side of (5.16) is AC-optimal. In this case, we may replace the ACOE by an *average cost optimality inequality* (ACOI), viz., (5.16).

(b) If, for each $i \in \mathbf{S}$, $U(i)$ is finite, then, in the above proof, $f_{\beta_n}(i) = f(i)$ for large n . Then we can write, for large n ,

$$\rho + h(i) = c(i, f(i)) + \beta_n \sum_j P(j \mid i, f(i)) h_{\beta_n}(j).$$

By Fatou’s lemma,

$$\rho + h(i) \geq c(i, f(i)) + \sum_j P(j \mid i, f(i)) h(j).$$

Consider the stronger assumption, below.

Assumption 5.17. Assumption 5.16 holds, and $\sum_j P(j \mid i, a) M(j) < \infty$, for all $a \in \mathbf{A}$ and $i \in \mathbf{S}$.

Under Assumption 5.17, using dominated convergence, it is easy to see that

$$\rho + h(i) = \min_{a \in U(i)} \left\{ c(i, a) + \sum_j P(j \mid i, a) h(j) \right\},$$

and we obtain the ACOE. If, for each $i \in \mathcal{S}$, there is a finite set $R_i \subset \mathcal{S}$ such that $P(j \mid i, \cdot) = 0$ for $j \notin R_i$, then Assumption 5.17 will obviously hold. Such a condition is satisfied for systems whose dynamics have a nearest-neighbour motion property [28].

(c) If there exists an $f \in \Pi_{SD}$, under which the process is ergodic, irreducible with an invariant measure $\eta(f) \in \mathcal{P}(\mathcal{S})$, and $\sum_i c(i, f(i))\eta(f)(i) < \infty$, then Assumptions 5.14 and 5.16 hold. Assumption 5.15 holds if $J_\beta^*(i)$ is increasing in i . Direct conditions implying Assumptions 5.14–5.17 can be found in [28], [32], [34], [76], [77], [155], [156], [172], [190]. See also [141], [142].

(d) Let $f \in \Pi_{SD}$ be a policy that attains the minimum on the right-hand side of (5.16). Fix an $i \in \mathcal{S}$. If the chain under f is positive recurrent at i , then we can show that equality holds at i in (5.16). However, the lack of positive recurrence at i may lead to strict inequality in (5.16). Cavazos-Cadena [33] had exhibited an example to demonstrate this. He has further shown in his example [33] that Assumptions 5.14–5.16 are satisfied, but the ACOE does not admit any solution.

5.3. The convex analytic approach. We will now describe Borkar's convex analytic approach for the average cost case [20]–[26]. The convex analytic approach to the AC-problem is a natural extension of the linear programming approach when the state/action spaces are no longer finite. In this approach, we view the control problem as the problem of minimizing a linear functional on the convex set of “ergodic occupation measures,” to be defined shortly [20]–[26]. This approach can also be used to treat other standard cost criteria, but it may be more involved for treating cases such as the DC criterion. On the other hand, it is more flexible and powerful for certain other purposes, e.g., pathwise average cost, constrained optimization problem, among others. Since the techniques involved here are entirely different from what we have thus far followed, we will embark on a more detailed discussion.

By replacing each $U(i)$ with $\prod_k U(k)$ and $P(j \mid i, \cdot)$ by its composition with the projection $\prod_k U(k) \rightarrow U(i)$, we may and will assume that the $U(i)$'s are replicas of a fixed compact metric space \mathbf{A} . We say that an $f \in \Pi_{SR}$ is *stable* if the corresponding process is positive recurrent. We will assume that, under an $f \in \Pi_{SR}$, the process has \mathcal{S} as its single communicating class. (This can be relaxed in some cases; see [26] for a discussion on this.) Therefore, f will have a unique invariant measure $\eta(f) \in \mathcal{P}(\mathcal{S})$ satisfying

$$\eta(f)P(f) = \eta(f).$$

Let Π_{SSR} denote the space of stable stationary policies. Π_{SSD} is defined analogously. For an $f \in \Pi_{SSR}$, denote by $\hat{\eta}(f) \in \mathcal{P}(\mathcal{S} \times \mathbf{A})$ the “ergodic occupation measure” defined by

$$\int_{\mathcal{S} \times \mathbf{A}} g d\hat{\eta}(f) = \sum_{i \in \mathcal{S}} \eta(f)(i) \int_{\mathbf{A}} g(i, a) f(i)(da)$$

for $g \in C_b(\mathcal{S} \times \mathbf{A})$. We will consider the sample path average cost optimality, which is stronger than the usual AC-optimality. Let

$$I_R = \{\hat{\eta}(f) : f \in \Pi_{SSR}\}, \quad I_D = \{\hat{\eta}(f) : f \in \Pi_{SSD}\}.$$

Note that $\hat{\eta}(f)$ can only be defined for an $f \in \Pi_{SSR}$. To consider optimality in Π , we will need to consider the following empirical processes. Let $\pi \in \Pi$ and let (X_t, A_t) be the corresponding processes with initial law $\mu \in \mathcal{P}(\mathcal{S})$. Define the $\mathcal{P}(\mathcal{S} \times \mathbf{A})$ -valued empirical process $\{\nu_t\}_{t \geq 1}$ by

$$(5.17) \quad \nu_t(C \times D) = \frac{1}{t} \sum_{s=0}^{t-1} I\{X_s \in C, A_s \in D\}, \quad t \geq 1,$$

for C, D Borel in \mathcal{S}, \mathcal{A} , respectively. Let $\overline{\mathcal{S}} = \mathcal{S} \cup \{\infty\}$ be the one-point compactification of \mathcal{S} . By abuse of notation, we may identify ν_t with the element of $\mathcal{P}(\overline{\mathcal{S}} \times \mathcal{A})$ that restricts to it on $\mathcal{S} \times \mathcal{A}$. Since $\mathcal{P}(\overline{\mathcal{S}} \times \mathcal{A})$ is compact, $\{\nu_t\}$, viewed as a sequence of $\mathcal{P}(\overline{\mathcal{S}} \times \mathcal{A})$ -valued random variables, converges to a sample path dependent compact limit set in $\mathcal{P}(\overline{\mathcal{S}} \times \mathcal{A})$. We characterize this set in Lemma 5.1, below, the statement of which calls for some new notation. Note that any element $\nu \in \mathcal{P}(\overline{\mathcal{S}} \times \mathcal{A})$ can be decomposed as

$$(5.18) \quad \nu(B) = \delta_\nu \nu'(B \cap (\mathcal{S} \times \mathcal{A})) + (1 - \delta_\nu) \nu''(B \cap (\{\infty\} \times \mathcal{A}))$$

for B Borel in $\overline{\mathcal{S}} \times \mathcal{A}$, $\delta_\nu \in [0, 1]$ is uniquely specified and $\nu' \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ (respectively, $\nu'' \in \mathcal{P}(\{\infty\} \times \mathcal{A})$) is uniquely specified if $\delta_\nu > 0$ (respectively, $\delta_\nu < 1$). We may render ν', ν'' unique at all times by imposing an arbitrary fixed choice thereof when $\delta_\nu = 0$, respectively, 1.

LEMMA 5.1. *Outside a set of zero probability (with respect to \mathcal{P}_μ^π), the following holds: For any limit point ν of $\{\nu_t\}$ in $\mathcal{P}(\overline{\mathcal{S}} \times \mathcal{A})$ for which $\delta_\nu > 0$,*

$$\nu' = \hat{\eta}(f)$$

for some $f \in \Pi_{SSR}$.

Proof. By the martingale stability theorem [117, p. 53],

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{s=1}^t \left[I\{X_s = i\} - E_\mu^\pi \left[I\{X_s = i\} \mid \mathfrak{F}_{s-1} \right] \right] \\ = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{s=1}^t \left[I\{X_s = i\} - \sum_{j \in \mathcal{S}} P(i \mid j, A_{s-1}) I\{X_{s-1} = j\} \right] \\ = \lim_{t \rightarrow \infty} \left[\nu_t(\{i\} \times \mathcal{A}) - \int P(i \mid \cdot, \cdot) d\nu_t \right] \\ = 0 \quad \text{a.s.,} \end{aligned}$$

for each $i \in \mathcal{S}$. Consider a sample path outside the set of zero probability on which the above fails for any $i \in \mathcal{S}$. Then, for any ν as in the statement of the lemma, we must have

$$\nu'(\{i\} \times \mathcal{A}) \geq \int P(i \mid \cdot, \cdot) d\nu', \quad i \in \mathcal{S}.$$

Note that an inequality is obtained here, since the second term on the right-hand side of (5.18) is obviously nonnegative. Summing over $i \in \mathcal{S}$ on both sides, it follows that equality must hold. Decomposing ν' as $\nu'(i, da) = \bar{\nu}(i) f(i)(da)$, where $\bar{\nu} \in \mathcal{P}(\mathcal{S})$ is the marginal on \mathcal{S} and $i \mapsto f(i) \in \mathcal{P}(\mathcal{A})$ is a version of the regular conditional law that defines an element of Π_{SR} , we obtain

$$\bar{\nu}(i) = \sum_{j \in \mathcal{S}} \bar{\nu}(j) P(f)(i, j).$$

Hence, $\bar{\nu} = \eta(f)$, and the conclusion follows. \square

LEMMA 5.2. *The sets I_R and I_D are closed; also, I_R is convex and has its extreme points in I_D .*

Proof. Let $\hat{\eta}(f_n) \in I_R$ and $\hat{\eta}(f_n) \rightarrow \nu$ for some ν in $\mathcal{P}(\mathcal{S} \times \mathcal{A})$. Then, for all $i \in \mathcal{S}$,

$$\hat{\eta}(f_n)(\{i\} \times \mathcal{A}) = \int P(i \mid \cdot, \cdot) d\hat{\eta}(f_n), \quad n \geq 1.$$

Letting $n \rightarrow \infty$, $\nu(\{i\} \times \mathcal{A}) = \int P(i \mid \cdot, \cdot) d\nu$. Now argue as in the proof of the preceding lemma to conclude that $\nu = \hat{\eta}(f)$ for some $f \in \Pi_{SSR}$. This proves that I_R is closed. The proof that I_D is closed is similar. Let $f_1, f_2 \in \Pi_{SSR}$ and $0 \leq \lambda \leq 1$. Define $f \in \Pi_{SSR}$ as follows:

$$f(i) = \frac{\lambda \eta(f_1)(i) f_1(i) + (1 - \lambda) \eta(f_2)(i) f_2(i)}{\lambda \eta(f_1)(i) + (1 - \lambda) \eta(f_2)(i)}.$$

Then using the properties of invariant measures, it is not difficult to see that

$$\begin{aligned} \eta(f) &= \lambda \eta(f_1) + (1 - \lambda) \eta(f_2), \\ \hat{\eta}(f) &= \lambda \hat{\eta}(f_1) + (1 - \lambda) \hat{\eta}(f_2), \end{aligned}$$

showing that I_R is convex. Now let $f \in \Pi_{SSR}$ be such that, for some $i_0 \in \mathcal{S}$ and $0 < \lambda < 1$, there exist $\phi_1, \phi_2 \in \mathcal{P}(\mathcal{A})$ such that

$$\begin{aligned} \int P(\cdot \mid i_0, a) f(i_0)(da) &= \lambda \int P(\cdot \mid i_0, a) \phi_1(da) + (1 - \lambda) \int P(\cdot \mid i_0, a) \phi_2(da), \\ \int P(\cdot \mid i_0, a) \phi_1(da) &\neq \int P(\cdot \mid i_0, a) \phi_2(da). \end{aligned}$$

Define $f_1, f_2 \in \Pi_{SSR}$ as

$$f_i(j) = \begin{cases} f(j), & j \neq i_0, \\ \phi_i, & j = i_0. \end{cases}$$

Then it can be shown [24] that $f_1, f_2 \in \Pi_{SSR}$, and any two of $\eta(f)$, $\eta(f_1)$, $\eta(f_2)$ are distinct from each other. Let $b \in (0, 1)$ be such that

$$\lambda = b\eta(f_1)(i_0) / (b\eta(f_1)(i_0) + (1 - b)\eta(f_2)(i_0)).$$

Then we can argue as before to conclude that $\hat{\eta}(f) = b\hat{\eta}(f_1) + (1 - b)\hat{\eta}(f_2)$. Therefore $\hat{\eta}(f)$ is not an extreme point of I_R . This implies that, for $\hat{\eta}(f')$ to be an extreme point of I_R , $P(\cdot \mid i, a)$ must be constant over $a \in \text{supp}(f'(i))$, for each $i \in \mathcal{S}$. Hence, $P(f'') = P(f')$, for all $f'' \in \Pi_{SSR}$ such that $\text{supp}(f''(i)) \subset \text{supp}(f'(i))$, for each $i \in \mathcal{S}$. In this case, $\eta(f'') = \eta(f')$. Suppose that for some i , say $i = 1$, there exist $\alpha \in (0, 1)$ and $\phi'_1, \phi'_2 \in \mathcal{P}(\mathcal{A})$, $\phi'_1 \neq \phi'_2$, such that $f'(1) = \alpha\phi'_1 + (1 - \alpha)\phi'_2$. Define $f'_1, f'_2 \in \Pi_{SSR}$ by

$$f'_k = \begin{cases} \phi'_k & \text{if } i = 1, \\ f'(i) & \text{if } i \neq 1, \end{cases} \quad k = 1, 2.$$

It follows that $\eta(f') = \eta(f'_1) = \eta(f'_2)$. It is also easy to check that

$$\begin{aligned} \hat{\eta}(f') &= \alpha\hat{\eta}(f'_1) + (1 - \alpha)\hat{\eta}(f'_2), \\ \hat{\eta}(f'_1) &\neq \hat{\eta}(f'_2), \end{aligned}$$

which contradicts the extremality of $\hat{\eta}(f')$. Hence, $f'(1)$ must be a Dirac measure. Applying this argument to each $i \in \mathbf{S}$, we deduce that $f \in \Pi_{SSD}$. From this, it follows that the extreme points of I_R lie in I_D . \square

We now proceed to show the existence of a sample path average cost optimal $f \in \Pi_{SSD}$. It is clear that a blanket stability condition or some condition on the cost that penalizes unstable behavior is required to give the desired existence. For example, consider the case where $c(i, a) = \exp(-i)$, which rewards unstable behavior. Clearly, the cost for any $f \in \Pi_{SSR}$ is almost surely positive. On the other hand, provided that $\Pi_{SSR} \neq \Pi_{SR}$, there exists an unstable policy in Π_{SR} that results in an almost-sure zero cost and is, therefore, optimal (the hypothesis that under some $f \in \Pi_{SR}$ the process has \mathbf{S} as its single communicating class plays a crucial role in this assertion). We want to rule out this possibility, as stability is a very desirable property of a policy. We wish to find conditions under which our goal will be achieved. Let $f \in \Pi_{SSR}$. Define

$$\rho(f) := \int c d\hat{\eta}(f), \quad \rho^* := \inf_{f \in \Pi_{SSR}} \rho(f).$$

Note that, under $f \in \Pi_{SSR}$, $J(i, f) = \rho(f)$ for each $i \in \mathbf{S}$. We consider two sets of hypotheses.

Assumption 5.18 (the near-monotonicity condition). It holds that

$$\liminf_{i \rightarrow \infty} \min_{a \in \mathbf{A}} c(i, a) > \rho^*.$$

Intuitively, Assumption 5.18 penalizes the drift of the process away from some finite set, requiring the optimal policy to exert some kind of a “centripetal force” pushing the process back toward this finite set. Thus, the optimal policy gains the desired stability property. If $c(i, a) = k(i)$ for some $k : \mathbf{S} \rightarrow \mathbb{R}_+$ and $k(i)$ is increasing, then this condition will automatically be satisfied. Such penalizing conditions quite often occur in queueing applications (see [20], [155], [156], [172], [190]).

Assumption 5.19 (stability condition (cf. Assumptions 5.7–5.12)). $\Pi_{SR} = \Pi_{SSR}$ and I_R is compact.

Assumption 5.19'. Equivalent conditions to Assumption 5.19 are

- (i) $\Pi_{SD} = \Pi_{SSD}$ and I_D is compact;
- (ii) The mean return times to a prescribed state (say 0) are uniformly integrable over all $f \in \Pi_{SR}$;
- (iii) This is the same as (ii), but with Π_{SD} replacing Π_{SR} .

THEOREM 5.10. *Under Assumption 5.18 or Assumption 5.19, there exists an $f \in \Pi_{SSD}$, which is sample path average cost optimal in Π_{SR} .*

Proof. From Lemma 5.2, it can be shown by an application of Choquet’s theorem [25], [26] that, if $\nu \mapsto \int c d\nu$ attains its minimum on I_R , it will do so for an $f \in \Pi_{SD}$. Under Assumption 5.19, it can be shown that $f \mapsto \hat{\eta}(f)$ is continuous. Therefore, the desired result follows under Assumption 5.19. We next consider the case under Assumption 5.18. Let $f_n \in \Pi_{SR}$ be such that $\rho(f_n) \downarrow \rho^*$. By identifying $\hat{\eta}(f_n)$ with the element of $\mathcal{P}(\bar{\mathbf{S}} \times \mathbf{A})$ that restricts to it on $\mathbf{S} \times \mathbf{A}$ for each n and then dropping to a subsequence if necessary, we may assume that $\hat{\eta}(f_n) \rightarrow \nu$ in $\mathcal{P}(\bar{\mathbf{S}} \times \mathbf{A})$ for some ν . Let $n \rightarrow \infty$ in the equation

$$\hat{\eta}(f_n)(\{j\} \times \mathbf{A}) = \int P(j \mid \cdot, \cdot) d\hat{\eta}(f_n), \quad j \in \mathbf{S}$$

and argue as in Lemma 5.1 to conclude that, for ν' as in (5.18), $\delta_\nu > 0$ implies that

$$\nu'(\{j\} \times \mathbf{A}) = \int P(j \mid \cdot, \cdot) d\nu', \quad j \in \mathbf{S}.$$

Decomposing ν' as $\nu'(i, da) = \bar{\nu}(i)f(i)(da)$, $i \in \mathbf{S}$, we have $\bar{\nu} = \eta(f)$ and therefore $\nu' = \hat{\eta}(f)$. Let $c_m = c \wedge m$ for $m \geq 1$ and pick $\varepsilon > 0$ such that Assumption 5.18 continues to hold with $\rho^* + \varepsilon$ in place of ρ^* . Then

$$\begin{aligned} \rho^* &= \lim_{n \rightarrow \infty} \int c d\hat{\eta}(f_n) \\ &\geq \lim_{n \rightarrow \infty} \int c_m d\hat{\eta}(f_n) \\ &\geq \delta_\nu \int c_m d\hat{\eta}(f) + (1 - \delta_\nu)((\rho^* + \varepsilon) \wedge m). \end{aligned}$$

Letting $m \rightarrow \infty$,

$$\rho^* \geq \delta_\nu \rho^* + (1 - \delta_\nu)(\rho^* + \varepsilon).$$

This is possible only if $\delta_\nu = 1$ and $\int c d\hat{\eta}(f) = \rho^*$. \square

The above theorem, however, does not ensure optimality of the cost-minimizing policy in I_R with respect to arbitrary policies. For the near-monotone case, this can be resolved without any further assumptions, but, for the stable case, we need the following.

Assumption 5.20. If $\tau = \min\{t \geq 1 : X_t = 0\}$, then

$$\sup_{\pi \in \Pi} E_0^\pi[\tau^2] < \infty.$$

Remark 5.9. Assumption 5.20 clearly implies Assumption 5.19. The converse need not be true, as can be shown by an explicit example [24]. Some sufficient conditions for Assumption 5.20 are (i) a Lyapunov condition [28], which we will describe shortly (cf. Theorem 5.11), (ii) the strong uniform recurrence condition of Doeblin and its variants [178], and (iii) the condition that there exist an $N < \infty$ for which

$$\sup_{\pi \in \Pi} \sup_i \mathcal{P}_i^\pi(\tau \geq N) < 1,$$

where τ is as above.

THEOREM 5.11. *Under Assumption 5.18 or Assumption 5.20, there exists an $f \in \Pi_{SD}$, which is sample path average cost optimal.*

Proof. Under Assumption 5.20, it can be shown [26] that the processes ν_t as defined in (5.17) are tight over Π . Therefore, δ_ν as in the statement of Lemma 5.1 may be taken to be 1. This resolves the case under Assumption 5.20. Under (A5.18), let ν be a limit point of $\{\nu_t\}$ in $\mathcal{P}(\bar{\mathbf{S}} \times \mathbf{A})$ along some subsequence. Then, as in the proof of Theorem 5.9, it can be shown that

$$(5.19) \quad \liminf_{t \rightarrow \infty} \int c d\nu_t \geq \rho^*.$$

Since this is true for any limit point ν of $\{\nu_t\}$ in $\mathcal{P}(\bar{\mathbf{S}} \times \mathbf{A})$ and for all sample points outside a set of probability zero, the desired result follows in this case also. \square

Remark 5.10. Some open problems arising in this context are:

(i) Can Assumption 5.20 be replaced by Assumption 5.19 while retaining the desired optimality?

(ii) If $\Pi_{SR} = \Pi_{SSR}$, will Assumption 5.19 hold automatically?

Remark 5.11. The condition in (5.19) implies a much stronger optimality, which will be discussed in §6.

Now, after the existence result of Theorem 5.11, an alternative treatment of the ACOE is possible. We will present a brief description without proofs. For details, see [24], [26], [28]. Define $h : \mathbf{S} \rightarrow \mathbb{R}$ by

$$(5.20) \quad h(i) = E_i^{f_0} \left[\sum_{t=0}^{\tau-1} (c(X_t, f_0(X_t)) - \rho^*) \right], \quad i \in \mathbf{S},$$

where $\tau = \min\{t \geq 1 : X_t = 0\}$ and $f_0 \in \Pi_{SD}$ is any sample path average cost optimal policy. In [22], [24], it is shown that $(h(\cdot), \rho^*)$ satisfies the ACOE under the following additional hypothesis called stability under local perturbations.

Assumption 5.21. Given an $f \in \Pi_{SSD}$ with $\rho(f) < \infty$, any $f' \in \Pi_{SD}$ obtained from f by changing the actions at most finitely many states is also stable and $\rho(f') < \infty$.

A sufficient, though not necessary, condition for Assumption 5.21 to hold is that every state has at most finitely many neighbors; i.e., for each $i \in \mathbf{S}$, there is a finite set $R_i \subset \mathbf{S}$ such that $P(j | i, \cdot) = 0$ for $j \notin R_i$.

In many cases, the solution (ρ^*, h) of the ACOE can be characterized (Theorem 5.12, below). The usual characterization of AC-optimal $f \in \Pi_{SD}$ in terms of the ACOE can also be proved for the foregoing.

THEOREM 5.12. *Assume Assumption 5.18 and let f_0, h be defined as above (cf. (5.20)). Let*

$$H = \{(\rho, w) : (\rho, w) \text{ satisfies the ACOE}, w(0) = 0, \inf w(\cdot) > -\infty\}.$$

Then (ρ^, h) is the unique element of H corresponding to the minimum value of ρ (i.e., if (ρ', w') is another element of H , then $\rho' \geq \rho^*$ with equality if and only if $w' = h$). Now, instead of Assumption 5.18, suppose that c is bounded and the following Lyapunov condition holds: There exists an $w : \mathbf{S} \rightarrow \mathbb{R}_+$, a finite $A \subset \mathbf{S}$ and an $\varepsilon > 0$ such that*

(a) $0 \in A$ and the set $\{i \in A^c : P(j | i, a) > 0, \text{ for some } j \in A, a \in \mathbf{A}\}$ is finite;

(b) $\lim_{i \rightarrow \infty} w(i) = \infty$;

(c) Under any $\pi \in \Pi, \mu \in \mathcal{P}(\mathbf{S})$

$$E_\mu^\pi [(w(X_{t+1}) - w(X_t) + \varepsilon) I\{X_t \notin A\} | \mathfrak{F}_t] \leq 0, \quad a.s.;$$

(d) There exists a random variable Z and a scalar $\lambda > 0$ such that $E[\exp(\lambda Z)] < \infty$ and, for all $b \geq 0$,

$$\mathcal{P}_\mu^\pi (|w(X_{t+1}) - w(X_t)| > b | \mathfrak{F}_t) \leq P(Z > b).$$

Then (ρ^, h) is the unique solution of the ACOE in the class $\{(\rho, w) : w(0) = 0, \limsup_{i \rightarrow \infty} h(i)/w(i) < \infty\}$.*

Remark 5.12. An alternative “intrinsic” formulation of the ACOE is also possible. For any $f \in \Pi_{SSD}$, define $h_f : \mathbf{S} \rightarrow \mathbb{R}$ by

$$h_f(i) = E_i^f \left[\sum_{t=0}^{\tau-1} (c(X_t, f(X_t)) - \rho(f)) \right], \quad i \in \mathbf{S}.$$

We say that f is *locally AC-optimal* if it yields a lower cost than any other element of Π_{SD} obtainable from f by changing f in at most finitely many states. In addition to the foregoing hypotheses, assume that every locally AC-optimal f is AC-optimal (for bounded c , a sufficient condition for this is that $\Pi_{SD} = \Pi_{SSD}$ and $\{\eta(f) : f \in \Pi_{SSD}\}$ is tight). We then have that f is sample path average cost optimal if and only if, for $i \in \mathbf{S}$,

$$h_f(i) = \inf_a \left\{ \sum_j P(j | i, a) h_f(j) + c(i, a) - \rho(f) \right\}.$$

This statement is “intrinsic” in the sense that all quantities (i.e., $h_f, \rho(f)$) are computable in terms of f . An interesting open problem is to characterize the most general conditions under which local AC-optimality implies AC-optimality.

Remark 5.13. The Lyapunov condition in Theorem 5.12(ii) implies Assumption 5.20 and has many other implications [26], but condition (ii)(d) there is rather strong, and, due to this, it may be difficult to construct such a function in a given situation. A partial answer to this question is given in [74]. It would be interesting to investigate if the Lyapunov conditions studied by [55], [91] (cf. Assumption 5.13), which do not involve condition (ii)(d) above, imply Assumption 5.20.

6. Borel state and action spaces. We consider in this section the case in which \mathbf{S} and \mathbf{A} are general Borel spaces. This is a natural setting for many problems, e.g., control of stock in water reservoirs, allocation of a resource between production and consumption, control of biological populations, harvesting a natural resource; see [17], [51], [82], and references therein for several examples. Also, the equivalent formulation of POCMP in terms of the conditional distribution of the (unobservable) state leads to a problem with an uncountable Borel state space, as we see in §7.

In this more general context, the ACOE is written as

$$(6.1) \quad \begin{aligned} \rho(x) + h(x) &= \inf_{a \in U(x)} \left\{ c(x, a) + \int_{\mathbf{S}} h(y) P(dy | x, a) \right\} \\ &= T(h)(x), \quad x \in \mathbf{S}, \end{aligned}$$

where $\rho, h \in \mathcal{M}(\mathbf{S})$. As in §5, a pair of functions (ρ, h) as above is called a solution to the ACOE, and, if ρ and h are bounded, we will say that the solution is bounded. Also, as in Theorem 5.1, our aim is to relate the AC problem to the existence of solutions to the ACOE. We have the following theorem.

THEOREM 6.1. *Suppose that (ρ, h) is a solution to the ACOE and that, for each policy $\pi \in \Pi_M$, the following holds:*

$$(6.2) \quad \lim_{t \rightarrow \infty} E_x^\pi \left[\frac{h(X_t)}{t} \right] = 0 \quad \forall x \in \mathbf{S}.$$

Then we have the following:

(i) *There holds*

$$(6.3) \quad \limsup_{n \rightarrow \infty} \frac{1}{n+1} E_x^\pi \left[\sum_{t=0}^n \rho(X_t) \right] \leq J(x, \pi),$$

and if $\pi \in \Pi_{SD}$ is such that $\pi(x)$ attains the infimum in (6.1), then equality is attained in (6.3);

(ii) *If $\rho(x) = \rho^* \in \mathbb{R}$, for all $x \in \mathcal{S}$, then $J^*(x) = \rho^*$, for all $x \in \mathcal{S}$, and any $\pi^* \in \Pi_{SD}$ such that $\pi^*(x)$ attains the infimum in (6.1) is average optimal.*

The proof of Theorem 6.1 follows that of Theorem 5.1 and is essentially contained in [177], more explicitly in [80], [148]; see also [78, pp. 66–68], [82, pp. 53–55], [150, pp. 93–94]. Note that (i), above, says that if $\rho(\cdot)$ is taken as the cost function to define another CMP $(\mathcal{S}, \mathcal{A}, U, P, \rho)$ then, for any $\pi \in \Pi_M$, the average cost incurred under the cost function $\rho(\cdot)$ does not exceed that under cost function $c(\cdot, \cdot)$.

Given the results above, it is of interest to find conditions under which there exists a solution (ρ, h) to the ACOE, satisfying (6.2). If h is bounded, then (6.2) is satisfied trivially. Also, if the random variables $\{h(X_t)\}$ are uniformly integrable under \mathcal{P}_x^π , for $\pi \in \Pi_M$ and $x \in \mathcal{S}$, then there exists a constant $0 < K_x^\pi < \infty$ such that $E_x^\pi[|h(X_t)|] \leq K_x^\pi$. Hence, if such a uniform integrability condition holds under *every* $\pi \in \Pi_M$ and $x \in \mathcal{S}$, then (6.2) is also satisfied trivially. The latter approach has been used by Shwartz and Makowski for some queueing problems [166]–[168].

6.1. Bounded costs. We first assume that $c(\cdot, \cdot)$ is bounded. When there are bounded solutions (ρ, h) to the ACOE, then much stronger results than those in Theorem 6.1 (i) can be obtained. To state these, some definitions are needed.

Let R and H be bounded, measurable, real-valued functions on \mathcal{S} , i.e., $R, H \in \mathcal{M}_b(\mathcal{S})$ and let $\pi^* \in \Pi$. Following the terminology of Dynkin and Yushkevich [51], the triplet (R, H, π^*) is said to be *canonical* if

$$(6.4) \quad J_N(x, \pi^*, H) = J_N^*(x, H) = H(x) + NR(x) \quad \forall N \in \mathbb{N}_0, \quad x \in \mathcal{S},$$

and $\pi^* \in \Pi$ is said to be a *canonical policy* if it is an element of some canonical triplet. Note that, if (R, H, π^*) is a canonical triplet, then π^* is N -stage optimal, for all $N \in \mathbb{N}_0$, when H is taken as the terminal cost. This concept was introduced by Yushkevich [204]. For finite models, Denardo and Fox [37] used a similar approach.

A policy $\pi^* \in \Pi$ is said to be *strong average optimal* if

$$(6.5) \quad \limsup_{N \rightarrow \infty} \frac{1}{N} J_N(x, \pi^*) \leq \liminf_{N \rightarrow \infty} \frac{1}{N} J_N(x, \pi) \quad \forall x \in \mathcal{S}, \pi \in \Pi.$$

Alternate definitions of strong average optimality are given in [69], [70]. Clearly, a strong average optimal policy π^* is also average optimal, and the limit of the sequence $\{1/N J_N(x, \pi^*)\}$, as $N \rightarrow \infty$, exists. An interpretation of (6.5) is that the “most pessimistic” average performance under π^* is no worse than the most “optimistic” performance under any other policy. We have the following result.

THEOREM 6.2. *Let $\pi^* \in \Pi_{SD}$, let $\rho, h \in \mathcal{M}_b(\mathcal{S})$, and let $c \in \mathcal{M}_b(\mathcal{K})$. Then (ρ, h, π^*) is a canonical triplet if and only if*

$$(6.6) \quad \rho(x) = \inf_{a \in U(x)} \left\{ \int_{\mathcal{S}} \rho(y) P(dy \mid x, a) \right\}$$

and

$$(6.7) \quad \rho(x) + h(x) = \inf_{a \in U(x)} \left\{ c(x, a) + \int_{\mathcal{S}} h(y) P(dy \mid x, a) \right\}$$

and $\pi^*(x)$ attains the infimum in both (6.6) and (6.7), for all $x \in \mathcal{S}$.

Proof. Necessity. Let (ρ, h, π^*) be a canonical triplet. Then, by (6.4),

$$(6.8) \quad \begin{aligned} h(x) + \rho(x) + N\rho(x) &= J_{N+1}^*(x, h) \\ &= T(J_N^*)(x) \\ &= c(x, \pi^*(x)) + \int_{\mathcal{S}} J_N^*(y, h) P(dy \mid x, \pi^*(x)). \end{aligned}$$

Since $J_0(x, \pi^*, h) = J_0^*(x, h) = h(x)$, then (6.7) follows from (6.8) by letting $N = 0$. Furthermore, since $\rho(\cdot)$, $h(\cdot)$, and $c(\cdot, \cdot)$ are bounded, then dividing both sides of (6.8) by N and letting $N \rightarrow \infty$ yields (6.6).

Sufficiency. Let (ρ, h) satisfy (6.6) and (6.7) and let $\pi^*(x)$ attain the infimum in these expressions. We use induction to show that (ρ, h, π^*) is a canonical triplet. For $N = 0$, this is trivially satisfied. Suppose that $N \in \mathbb{N}_0$ is the first integer for which (6.4) fails; then

$$\begin{aligned} J_N^*(x, h) &= T(J_{N-1}^*)(x) \\ &= T(h + (N-1)\rho)(x) \\ &= \inf_{a \in U(x)} \left\{ c(x, a) + \int_{\mathcal{S}} h(y) P(dy \mid x, a) + (N-1) \int_{\mathcal{S}} \rho(y) P(dy \mid x, a) \right\} \\ &\geq T(h)(x) + (N-1) \inf_{a \in U(x)} \left\{ \int_{\mathcal{S}} \rho(y) P(dy \mid x, a) \right\} \\ &= T(h)(x) + (N-1)\rho(x) = h(x) + N\rho(x). \end{aligned}$$

On the other hand,

$$\begin{aligned} J_N^*(x, h) &\leq J_N(x, \pi^*, h) \\ &= c(x, \pi^*(x)) + \int_{\mathcal{S}} J_{N-1}^*(y, \pi^*, h) P(dy \mid x, \pi^*(x)) \\ &= c(x, \pi^*(x)) + \int_{\mathcal{S}} [h(y) + (N-1)\rho(y)] P(dy \mid x, \pi^*(x)) \\ &= T(h)(x) + (N-1)\rho(x) = h(x) + N\rho(x) \end{aligned}$$

contradicting our hypothesis. Therefore, (ρ, h, π^*) is a canonical triplet. \square

The results in Theorem 6.2 were obtained by Yushkevich [204]; see also [51]. Note that (6.7) is the ACOE and that (6.6) allows $\rho(\cdot)$ to be treated as a constant, with respect to the optimization problem. Of course, if $\rho(x) = \rho^*$ for all $x \in \mathcal{S}$, then (6.6) is satisfied trivially. The coupled equations (6.6) and (6.7) were apparently introduced by Howard [95, pp. 61–62], in the context of finite state CMP for which, under some policies, $\{X_t\}$ has several ergodic classes, i.e., the so-called multichain case. In this case, different ergodic classes may have different optimal average cost, and $\rho(\cdot)$ gives this cost, as will be shown.

From Theorem 6.2, we see that the canonical policy π^* is a measurable selector for both (6.6) and (6.7). However, Assumption 2.2 in §2 is not enough to guarantee the existence of selectors in either (6.6) or (6.7), since ρ and h are assumed to be bounded and measurable functions, but not necessarily lower semicontinuous. For this situation, the following condition is needed.

Assumption 6.1. The transition kernel $P(\cdot \mid x, a)$ is *strongly continuous* in (x, a) ; that is, $u \in \mathcal{M}_b(\mathbf{S})$ implies that $\int_{\mathbf{S}} u(y)P(dy \mid \cdot, \cdot) \in C_b(\mathbf{K})$.

It follows that under Assumptions 2.1, 2.3, 6.1, measurable selectors exist for each of (6.6) and (6.7), and $\pi^* \in \Pi_{SD}$ will be a canonical policy if and only if it is a selector for both (6.6) and (6.7). If (ρ, h, π^*) is a canonical triplet, then (ρ, h) solves the ACOE, and (6.2) is satisfied, since h is bounded. Consequently, the results of Theorem 6.1 follow. The next result presents other important implications.

THEOREM 6.3. *Let (ρ, h, π^*) be a canonical triplet, and let $c \in \mathcal{M}_b(\mathbf{K})$. Then, for each $x \in \mathbf{S}$,*

- (i) $J_N(x, \pi^*) \leq J_N(x, \pi) + \text{span}(h)$, for every $\pi \in \Pi$;
- (ii) π^* is strong average optimal;
- (iii) $J(x, \pi^*) = J^*(x) = \rho(x)$;
- (iv) $h^-(x) + \rho(x)/(1 - \beta) \leq J_\beta^*(x) \leq h^+(x) + \rho(x)/(1 - \beta)$;
- (v) If $\rho(x) = \rho^* \in \mathbb{R}$, for all $x \in \mathbf{S}$, then, for every $\pi \in \Pi$,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t) \geq \rho^*, \quad \mathcal{P}_x^\pi\text{-a.s.},$$

when $X_0 = x$, and $\{A_t\}$ is generated using the policy π . Furthermore,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t) = \rho^*, \quad \mathcal{P}_x^\pi\text{-a.s.},$$

if and only if

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=0}^{N-1} \Phi(X_t, A_t) = 0, \quad \mathcal{P}_x^\pi\text{-a.s.},$$

where $\Phi : \mathbf{K} \rightarrow \mathbb{R}$ is given by

$$\Phi(x, a) := c(x, a) + \int h(y)P(dy \mid x, a) - \rho^* - h(x);$$

- (vi) π^* is sample path average cost optimal.

Proof. To prove (i), note that, for all $\pi \in \Pi$,

$$\begin{aligned} J_N(x, \pi^*, h) &= E_x^{\pi^*} \left[\sum_{t=0}^{N-1} c(X_t, A_t) + h(X_N) \right] \\ &\leq E_x^\pi \left[\sum_{t=0}^{N-1} c(X_t, A_t) + h(X_N) \right] = J_N(x, \pi, h). \end{aligned}$$

Hence,

$$\begin{aligned} J_N(x, \pi^*) &\leq J_N(x, \pi) + E_x^\pi[h(X_N)] - E_x^{\pi^*}[h(X_N)] \\ &\leq J_N(x, \pi) + \text{span}(h) \quad \forall \pi \in \Pi. \end{aligned}$$

By the boundedness of $h(\cdot)$, we have that

$$\lim_{N \rightarrow \infty} \frac{1}{N} J_N(x, \pi^*, h) = \lim_{N \rightarrow \infty} \left[\frac{h(x) + N\rho(x)}{N} \right] = \rho(x).$$

Furthermore, since $J_N(x, \pi^*, h) = J_N(x, \pi^*) + E_x^{\pi^*}[h(X_N)]$, then

$$\rho(x) = \lim_{N \rightarrow \infty} \frac{1}{N} J_N(x, \pi^*),$$

and (ii)–(iii) follows from (i).

Next, since (ρ, h) solve the ACOE, then (ρ, h^-) and (ρ, h^+) are also solutions to the ACOE. Since $h^-(\cdot) \leq 0 \leq h^+(\cdot)$, then by Lemma 2.1 we have that $T(h^-) \leq T(\beta h^-) = T_\beta(h^-)$, and $T(h^+) \geq T(\beta h^+) = T_\beta(h^+)$. Then, (iv) follows by induction, using Theorem 2.1 (iv); see [64].

Turning our attention to (v) and (vi), observe that, due to (6.7), $\Phi(x, a) \geq 0$ for all $(x, a) \in \mathbf{K}$. Also, by the (Markov) property (2.3) in §2, we have that, for any $\pi \in \Pi$,

$$\Phi(X_t, A_t) = E_x^\pi \left[c(X_t, A_t) + h(X_{t+1}) - \rho^* - h(X_t) \mid H_t, A_t \right], \quad \mathcal{P}_x^\pi\text{-a.s.}$$

Let

$$Z_t := c(X_t, A_t) + h(X_{t+1}) - h(X_t) - \rho^* - \Phi(X_t, A_t)$$

and

$$M_N := \sum_{t=0}^{N-1} Z_t = \sum_{t=0}^{N-1} c(X_t, A_t) - N\rho^* + h(X_N) - h(X_0) - \sum_{t=0}^{N-1} \Phi(X_t, A_t).$$

Note that $\{Z_t\}$ is a $(\mathfrak{G}_t, \mathcal{P}_x^\pi)$ martingale difference, where $\mathfrak{G}_t := \sigma(H_{t+1}, A_{t+1})$. Since $\{Z_t\}$ is bounded uniformly in t , by the martingale stability theorem

$$\lim_{N \rightarrow \infty} \frac{M_N}{N} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=0}^{N-1} Z_t = 0, \quad \mathcal{P}_x^\pi\text{-a.s.}$$

Therefore, by the boundedness of $h(\cdot)$,

$$\lim_{N \rightarrow \infty} \left[\frac{1}{N} \sum_{t=0}^{N-1} c(X_t, A_t) - \rho^* - \frac{1}{N} \sum_{t=0}^{N-1} \Phi(X_t, A_t) \right] = 0, \quad \mathcal{P}_x^\pi\text{-a.s.}$$

Finally, (v) and (vi) follow, since $\Phi(x, a) \geq 0$ for all $(x, a) \in \mathbf{K}$ and since, for a canonical policy π^* , $\Phi(X_t, A_t) = 0$, $\mathcal{P}_x^{\pi^*}$ -a.s. \square

The results in (i)–(iii) of Theorem 6.3 are essentially contained in [51, Chap. 7]; that in (iv) is motivated by similar results in [136] and [64]; (v) and (vi) are due to Georgin [72], see also [82, pp. 52–55]. Also, the function Φ defined in (v) was introduced by Mandl [124] and is often referred to as *Mandl's discrepancy function*.

In view of Theorem 6.3, it follows that a canonical triplet yields the desired results. We therefore look for conditions on the primary objects like the cost function c and transition kernel P , which imply the existence of a canonical triplet, so that the theory

can be used in a given practical situation. To this end, a standard procedure is to assume some ergodicity conditions that will ensure the existence of a bounded solution to the ACOE. We have already discussed several such conditions for the countable state case (cf. Assumptions 5.1–5.5). Analogues of such assumptions are also available in the literature, an extensive survey of which appears in [86]. We will focus on a particular ergodicity condition that not only subsumes many such conditions but also facilitates easily implementable numerical schemes. Our presentation here follows essentially that in [82, Chap. 3].

Assumption 6.2. There exists a number $\alpha < 1$ such that

$$\sup_{k, k' \in K} \|P(\cdot | k) - P(\cdot | k')\|_{TV} \leq 2\alpha,$$

where $\|\cdot\|_{TV}$ denotes the total variation norm.

Example 6.1. Let $\mathbf{S} = \mathbb{R}$, $\mathbf{A} \subset \mathbb{R}$, a compact set. Consider the system

$$X_{t+1} = F(X_t, A_t) + G(X_t)W_t, \quad X_0 = x,$$

where $F : \mathbb{R} \times \mathbf{A} \rightarrow \mathbb{R}$, $G : \mathbb{R} \rightarrow \mathbb{R}$ are bounded, continuous and $G(\cdot) > 0$, and $\{W_t\}$ is a sequence of independent $N(0, 1)$ random variables ($N(a, b)$ stands for the Gaussian distribution with mean a and variance b). In this case, the transition kernel is given by

$$P(\cdot | x, a) = N(F(x, a), G^2(x)).$$

Using the assumed conditions on F, G we can show that Assumption 6.2 holds. We omit the details. An important consequence of Assumption 6.2 is given below; for a proof and further discussion, see [82, Chap. 3].

LEMMA 6.1. *Suppose that Assumption 6.2 holds. Then, for any $f \in \Pi_{SD}$, the corresponding process $\{X_t\}$ has a unique invariant measure $\eta(f) \in \mathcal{P}(\mathbf{S})$ satisfying*

$$(6.9) \quad \|P^t(\cdot | x, f(x)) - \eta(f)(\cdot)\|_{TV} \leq 2\alpha^t, \quad t = 0, 1, \dots,$$

where $P^t(\cdot | x, f(x))$ denotes the t -step transition probability measure under f with $X_0 = x$.

Remark 6.1. (a) Lemma 6.1 also holds for any $f \in \Pi_{SR}$.

(b) It follows from (6.9) that, for any $f \in \Pi_{SD}$, $P^t(\cdot | x, f(x))$ converges to $\eta(f)$ in total variation norm, uniformly in x , and at a geometric rate.

(c) It is clear that, for any $f \in \Pi_{SD}$,

$$J(\mu, f) = \int_{\mathbf{S}} c(x, f(x))\eta(f)(dx)$$

for any initial law μ .

(d) Compare (6.9) with Assumption 5.5. In view of Theorem 5.4, Assumption 6.2 may be viewed as a representative counterpart of Assumptions 5.1–5.5 for the general state space case.

We now introduce the concept of span-contraction.

DEFINITION 6.1. Let $T : \mathcal{M}_b(\mathbf{S}) \rightarrow \mathcal{M}_b(\mathbf{S})$. T is said to be a *span-contraction* if, for some $\gamma \in [0, 1)$,

$$\text{span}(Tu - Tv) \leq \gamma \text{span}(u - v) \quad \text{for all } u, v \in \mathcal{M}_b(\mathbf{S}).$$

Let \sim be the equivalence relation on $\mathcal{M}_b(\mathbf{S})$ defined by $u \sim v$ if and only if there exists some constant C such that $u(x) - v(x) = C$ for all $x \in \mathbf{S}$. Let $\widetilde{\mathcal{M}}_b(\mathbf{S}) = \mathcal{M}_b(\mathbf{S}) / \sim$, the quotient space, endowed with the quotient norm induced by the span seminorm. For $v \in \mathcal{M}_b(\mathbf{S})$, let \tilde{v} denote the corresponding element of $\widetilde{\mathcal{M}}_b(\mathbf{S})$ and $\tilde{T} : \widetilde{\mathcal{M}}_b(\mathbf{S}) \rightarrow \widetilde{\mathcal{M}}_b(\mathbf{S})$ be the canonically induced map, i.e., $\tilde{T}\tilde{v} = \widetilde{Tv}$, $v \in \mathcal{M}_b(\mathbf{S})$. It is easily seen that, if T is a span-contraction on $\mathcal{M}_b(\mathbf{S})$, then \tilde{T} is a contraction on $\widetilde{\mathcal{M}}_b(\mathbf{S})$ and therefore has a unique fixed point. In turn, it follows that the map T has a span-fixed point; i.e., there exists a $v^* \in \mathcal{M}_b(\mathbf{S})$ such that $\text{span}(Tv^* - v^*) = 0$ or, equivalently, $Tv^* - v^*$ is a constant. It also follows that any two span-fixed points of T must differ by a constant.

We now replace Assumption 2.3 with the following

Assumption 6.3. (i) The multifunction $U(x)$ is continuous; (ii) $c(\cdot, \cdot) \in C_b(\mathbf{K})$.

We have the following result; for a proof, see [82, Lemma 3.5].

LEMMA 6.2. *Under Assumptions 2.2, 6.2, and 6.3, the operator T defined in (2.5) maps $C_b(\mathbf{S})$ to $C_b(\mathbf{S})$ and is a span-contraction.*

COROLLARY 6.1. *Under Assumptions 2.2, 6.2, and 6.3, the ACOE has a bounded solution $(\rho^*, h^*) \in \mathbb{R} \times C_b(\mathbf{S})$.*

Proof. This follows from the fact that there exists a $h^* \in C_b(\mathbf{S})$ such that $\text{span}(Th^* - h^*) = 0$. Hence, $Th^* = h^* + \rho^*$ for some constant ρ^* . \square

Remark 6.2. (a) Assume Assumptions 6.2 and 6.3. Let $(\rho^*, h^*) \in \mathbb{R} \times C_b(\mathbf{S})$ be a solution to the ACOE and fix $x_0 \in \mathbf{S}$. Define $h(\cdot) = h^*(\cdot) - h^*(x_0)$. Then (ρ^*, h) is also a solution to the ACOE. By the span-contraction property of T , it is the unique solution in $\mathbb{R} \times C_b(\mathbf{S})$ satisfying $h(x_0) = 0$; i.e., if $(\rho', h') \in \mathbb{R} \times C_b(\mathbf{S})$ is any other solution of the ACOE in $\mathbb{R} \times C_b(\mathbf{S})$ such that $h'(x_0) = 0$, then $\rho' = \rho$ and $h' = h$.

(b) In view of the span-contraction property of the operator T , the value iteration scheme described in §4 can be extended to this case; for details, we refer to [82, Chap. 3].

(c) Note that Corollary 6.1 asserts the existence of a canonical triplet.

Remark 6.3. In §4 we have identified the duality between the linear programming formulation and the ACOE under the irreducibility assumption. This has been extended by Yamada [203] to the case when the state space \mathbf{S} is a compact subset of \mathbb{R}^n and the transition law has a density that satisfies a certain “positivity” condition. Hernández-Lerma, Hennet, and Lasserre [84] have further extended this result to the Borel state space setting under Assumption 6.2.

Kurano [105]–[107] has studied the problem for compact state and action spaces, under the hypothesis of Doeblin. Doeblin’s condition for the general state space can be described as follows.

Assumption 6.4. There exists a nontrivial finite measure μ on $(\mathbf{S}, \mathcal{B}(\mathbf{S}))$, a positive integer ℓ , and an $\varepsilon > 0$ such that

$$P^\ell(A \mid x, f(x)) \geq 1 - \varepsilon \quad \text{if } \mu(A) \geq \varepsilon,$$

for all $f \in \Pi_{SD}$ and $x \in \mathbf{S}$.

THEOREM 6.4. *Let the state and action spaces be compact and Assumptions 6.3 and 6.4 hold. Then there exist an $f \in \Pi_{SD}$ and a set $A \in \mathcal{B}(\mathbf{S})$ with $\mu(A) > \varepsilon$ such that $P(A \mid x, f(x)) = 1$ for all $x \in \mathbf{S}$, and f is optimal, provided that the initial law is supported on the set A .*

Furthermore, assume the following.

Assumption 6.5 (reachability). For any $x \in \mathcal{S}$ and $D \in \mathcal{B}(\mathcal{S})$ with $\mu(D) > \varepsilon$ (μ and ε as in Assumption 6.4, there exists a $\pi \in \Pi$ such that

$$P_x^\pi \left(\bigcup_{t=0}^{\infty} \{X_t \in D\} \right) = 1.$$

Assumption 6.6. One of the following two conditions is satisfied:

- (i) $\mu(\partial D) = 0$ if $\mu(D) > 0$, where ∂D denotes the boundary of D ;
- (ii) For each $D \in \mathcal{B}(\mathcal{S})$ with $\mu(D) > \varepsilon$, $P(D \mid x, a)$ is continuous in (x, a) .

THEOREM 6.5. *Under Assumptions 6.3–6.6 there exists an $f \in \Pi_{SD}$, which is optimal.*

Remark 6.4. (a) The proof of Theorem 6.3 exploits the idea involved in Lemma 5.1 of extracting a stationary randomized policy from a limit point of empirical processes. A novel idea in [105] is to remove the randomization by using the ergodic decomposition of Markov processes under Assumption 6.4. The compactness is used to ensure the tightness of the empirical processes under any policy. This can be dropped if the cost function has a penalizing condition or if there is a blanket stability of Lyapunov type. The details closely mimic the development at the end of §5.

(b) Wijngaard [201] has also obtained the existence of an optimal $f \in \Pi_{SD}$ under Doeblin's condition using an operator theoretic method.

We will now discuss the vanishing discount approach to obtain a bounded solution to the ACOE. For a fixed $x_0 \in S$, let $h_\beta(\cdot) = J_\beta^*(\cdot) - J_\beta^*(x_0)$ denote the differential discounted value function. For a general state space, the usual diagonalization procedure used on a countable state space is not amenable. Nevertheless, if $h_\beta(\cdot)$ is uniformly bounded and equicontinuous, then we can use a more subtle diagonalization involving the Arzela–Ascoli theorem to take the required limits and obtain a bounded solution to the ACOE. This was studied by Ross [148]. Following [17], [72], [73], we will discuss some sufficient conditions to obtain the required uniform boundedness and equicontinuity of $h_\beta(\cdot)$.

Assumption 6.7. For each $\beta \in (\beta', 1)$, for some $0 < \beta' < 1$, and $f_\beta \in \Pi_{SD}$, the corresponding state process has a unique invariant probability measure $\eta(f_\beta)$ such that

$$(6.10) \quad \sup_{\substack{x \in S \\ \beta \in (\beta', 1)}} \sum_{t=1}^{\infty} \|P^t(\cdot \mid x, f_\beta(x)) - \eta(f_\beta)(\cdot)\|_{TV} < \infty.$$

The following result is now easy to establish.

LEMMA 6.3. *Under Assumptions 6.1, 6.3, and 6.7, $h_\beta(\cdot) := J_\beta^*(\cdot) - J_\beta^*(x_0)$, $x_0 \in S$ fixed, is uniformly bounded, and is equicontinuous for $\beta \in (\beta', 1)$.*

COROLLARY 6.2. *Under Assumptions 6.1, 6.3, and 6.7, the ACOE has a solution (ρ^*, h) such that $h \in C_b(S)$.*

Remark 6.5. If Assumption 6.4 is satisfied and we further impose the condition that, for every $f \in \Pi_{SD}$, the corresponding state process has a single ergodic class, then (6.10) holds. In particular, if $P(dy \mid x, a)$ has a density $p(y, x, a)$, with respect to some σ -finite measure μ , and there exists a nonnegative measurable function p_0 satisfying $\int p_0(y) \mu(dy) > 0$ and $p(y, x, a) \geq p_0(y)$, for all (x, a) , then Assumption 6.4 holds and (6.10) can be easily verified. If $(x, a) \rightarrow p(y, x, a)$ is continuous, then by Scheffe's theorem, $p(\cdot \mid x, a)$ is strongly continuous in (x, a) .

6.2. Unbounded costs. We now drop the boundedness condition on the cost function and discuss some recent developments involving refinements and extensions of the vanishing discount approach. Since for unbounded costs the uniform boundedness of the differential discounted value function $h_\beta(\cdot)$ is rather unnatural, we attempt to extend the procedure of [155], [156] to the present case. To this end, we make the following analogues of Assumptions 5.14–5.16.

Assumption 6.8. There exists a nonnegative function $b \in \mathcal{M}(\mathbf{S})$, a constant $M \geq 0$, and a sequence $\{\beta_n\} \subset (0, 1)$, $\beta_n \uparrow 1$, such that for all $x \in \mathbf{S}$, (i) $-M \leq h_{\beta_n}(x) \leq b(x)$, and (ii) $\int_{\mathbf{S}} b(y)P(dy | x, a) < \infty$, for all $a \in U(x)$.

Assumption 6.9. There exists a policy π and an initial state \hat{x} such that $J(\hat{x}, \pi) < \infty$.

Assumption 6.10. There exists $\beta' \in (0, 1)$ such that $\sup_{\beta \in (\beta', 1)} \tilde{h}_\beta(x) < \infty$, where $\tilde{h}_\beta(x) = J_\beta^*(x) - \inf_{x \in \mathbf{S}} J_\beta^*(x)$.

Assumption 6.11. The transition kernel $P(\cdot | x, a)$ is strongly continuous in a , for each $x \in \mathbf{S}$.

Under Assumptions 6.8 and 6.11, defining $h(x) = \liminf_{n \rightarrow \infty} h_{\beta_n}(x)$, $x \in \mathbf{S}$, and using Fatou's lemma, we can show that there exists a constant ρ^* such that

$$(6.11) \quad \lim_{n' \rightarrow \infty} (1 - \beta_{n'})J_{\beta_{n'}}^*(x) = \rho^* \quad \text{for all } x \in \mathbf{S},$$

where $\beta_{n'} \uparrow 1$ is a subsequence of $\{\beta_n\}$, and

$$(6.12) \quad \rho^* + h(x) \geq \min_{a \in U(x)} \left\{ c(x, a) + \int_{\mathbf{S}} h(y)P(dy | x, a) \right\}, \quad x \in \mathbf{S},$$

which is the ACOI (see (5.16)) for this case. Similarly, under Assumptions 6.9–6.11, we can find a constant ρ^* such that, along a suitable sequence $\beta_n \in (\beta', 1)$, $\beta_n \uparrow 1$, $\lim_{n \rightarrow \infty} (1 - \beta_n) \inf_{x \in \mathbf{S}} J_{\beta_n}^*(x) = \rho^*$. Then, defining $h(x) = \liminf_{n \rightarrow \infty} \tilde{h}_{\beta_n}(x)$, we can deduce (6.12). Thus, we have the following result.

THEOREM 6.6. *Under Assumptions 6.8 and 6.11 or under Assumption 6.9–6.11, there exists a constant ρ^* and a function h , which is bounded below and satisfies (6.12). Any policy $\pi \in \Pi_{SD}$ realizing the minimum on the right-hand side of (6.12) is average optimal and ρ^* is the minimum average cost.*

Remark 6.6. For details, we refer to [83], [85], [140]. In the case of a countable state space, a number of sufficient conditions on the transition kernel and the cost function that enable us to verify Assumptions 5.14–5.16 are available, as mentioned in §5. This does not seem to be the case for a general Borel state space model, although several interesting examples have been studied in [83], [85], and [140]. Also, Assumption 6.11 is a very strong condition and will not, in general, be satisfied for the transition kernel of the equivalent problem for a partially observable model. Thus, this case needs further investigation. Finally, note that Assumption 6.10 may in principle be easier to verify than Assumption 6.8.

Remark 6.7. We note that Theorem 6.6 provides only an ACOI, and not the ACOE. In many situations, the discounted value function is convex (e.g., in linear systems with quadratic cost [14]), or concave (e.g., the separated problem in partially observable models). This class of problems has been used in [61] to obtain the ACOE under Assumptions 6.8, 6.11, and some additional assumptions.

7. Partially observable controlled Markov processes. Thus far, we have assumed that the complete history of the process H_t is available to the decision-maker, at each stage $t \in T$. However, in many situations, some components of the state process may not be directly available to the controller since, e.g., it may be impossible or too costly to measure these. Furthermore, due to imprecisions in the measuring devices, only noisy observations of the state may be available. When these situations arise, the problem is said to be a partially observable controlled Markov process. Here, we study POCMP with finite or countably infinite state and observation spaces, and finite or compact action set. A major portion of our exposition concentrates on the vanishing discount method, where we see that the particular structure of the POCMP can be employed to yield stronger results than those available for general Borel spaces. We also review Borkar's convex analytic approach, specialized to the partially observable case [26].

7.1. Models with partial state information. The model for this problem is essentially that in [51, Chap. 8] and is as follows. The state process is described by a pair $\{X_t, Y_t\}_{t \in T}$ taking values in a product of Borel spaces $\mathbf{X} \times \mathbf{Y}$. Only the second component $\{Y_t\}_{t \in T}$ of the state process is available for decision-making, and, reflecting this, \mathbf{Y} is called the *observation* or *message space*, and Y_t the *observation process*. With \mathbf{A} denoting the action space, the evolution of the system is governed by a measurable stochastic kernel P on $\mathbf{X} \times \mathbf{Y}$ given $\mathbf{X} \times \mathbf{Y} \times \mathbf{A}$.

Let $\mu \in \mathcal{P}(\mathbf{X} \times \mathbf{Y})$ be an initial distribution of the state. Decomposing (disintegrating) the measure μ , we have

$$\mu(dx, dy) = \bar{Q}_0(dy) \psi_0(dx | y),$$

where \bar{Q}_0 is the marginal of μ on \mathbf{Y} and ψ_0 is a version of the regular conditional law, defined \bar{Q}_0 almost surely; we pick any version from this equivalence class and keep it fixed thereafter. Note that knowledge of μ , since the value of Y_0 is available to the controller, implies that an a posteriori distribution ψ_0 (given $Y_0 = y$) for the unobserved initial state is introduced. We include ψ_0 into the *observed history* by letting

$$\bar{H}_0 := \mathcal{P}(\mathbf{X}) \times \mathbf{Y}, \quad \bar{H}_t := \bar{H}_{t-1} \times \mathbf{Y} \times \mathbf{A}, \quad t \in \mathbb{N}_0.$$

The set of admissible actions is specified by a strict, measurable, compact-valued multifunction $U : \mathbf{Y} \rightarrow \mathcal{B}(\mathbf{A})$. Hence, in this context, an admissible policy is a sequence $\pi = \{\pi_t\}_{t \in T}$ of Borel measurable stochastic kernels π_t on \mathbf{A} given \bar{H}_t satisfying, for all $t \in T$, the constraint

$$\pi_t(U(y_t) | \bar{h}_t) = 1 \quad \forall \bar{h}_t \in \bar{H}_t.$$

The set of all admissible policies is again denoted by Π .

Remark 7.1. In general, decisions take into account past and present information, not just the last observation. Note that the constraints on the actions cannot depend on the unobservable component X_t of the state. If this type of constraint must be included in the model, then it must be provided to the controller as an additional observation. Similarly, if the cost process $\{c(X_t, Y_t, A_t)\}$ is available to the controller, then it should also be regarded as an additional component in the observation process [51, p. 201].

Remark 7.2. Quite often, μ is specified as

$$\mu(dx, dy) = Q_0(dy | x) \mu_0(dx),$$

where $\mu_0 \in \mathcal{P}(\mathbf{X})$ is an initial distribution for X_0 , and Q_0 is a stochastic kernel on \mathbf{Y} given \mathbf{X} [15, Chap. 10], [82, Chap. 4].

With $\mu \in \mathcal{P}(\mathbf{X} \times \mathbf{Y})$ and an admissible policy π specified, there exists a unique probability measure \mathcal{P}_μ^π on $(\Omega, \mathcal{B}(\Omega))$, where $\Omega := (\mathbf{X} \times \mathbf{Y} \times \mathbf{A})^\infty$, defined by

$$\begin{aligned} \mathcal{P}_\mu^\pi(dx_0, dy_0, da_0, \dots, da_{t-1}, dx_t, dy_t) \\ = \mu(dx_0, dy_0) \pi_0(da_0 \mid \psi_0, y_0) P(dx_1, dy_1 \mid x_0, y_0, a_0) \cdots \\ \pi_{t-1}(da_{t-1} \mid \psi_0, y_0, a_0, \dots, y_{t-1}) P(dx_t, dy_t \mid x_{t-1}, y_{t-1}, a_{t-1}). \end{aligned}$$

7.2. Transformation into a completely observable model. A common approach in the analysis of a partially observable (PO) model is to construct a completely observable (CO) model, equivalent to the original one in the sense that corresponding policies have equal costs. The advantages in doing this are obvious, since the theory of CO problems is much better developed. However, the price usually paid is that the dimensionality of the new state space is substantially larger than that of the original one.

Such an equivalent CO problem can be obtained in many ways. The main idea is to specify an *information state process* that summarizes, at each time, all relevant information for decision-making. Clearly, $\bar{H}_t = (\psi_0, Y_0, A_0, \dots, A_{t-1}, Y_t)$ can be used as an information state process, but this leads to a nonstationary CO model, in which “growing memory” difficulties arise; see [15, Chap. 10]. We present here the more standard approach where the inferential knowledge of X_t is summarized using its conditional probability distribution, given the entire observed history up to time t . We first present the construction of the equivalent CO model for general Borel state spaces and then specialize to models with countable state space. Also, the following assumption will be in effect throughout this section.

Assumption 7.1. The transition kernel $P(\cdot \mid x, y, a)$ and the cost function $c(x, y, a)$ do not depend on y , and $U(y) = \mathbf{A}$ for all $y \in \mathbf{Y}$.

7.2.1. Borel state space. Given a PO model $(\mathbf{X} \times \mathbf{Y}, \mathbf{A}, U, P, c)$ satisfying Assumption 7.1, we construct a CO model $(\mathcal{P}(\mathbf{X}), \mathbf{A}, \tilde{U}, \mathcal{K}, \tilde{c})$ as follows. Let $\{\Psi_t, Y_t\}_{t \in T}$ and $\{\tilde{H}_t\}_{t \in T}$ denote the state process and the history spaces, respectively. The set of admissible actions is selected by letting $\tilde{U}(\psi) = \mathbf{A}$ for all $\psi \in \mathcal{P}(\mathbf{X})$. We define the cost function \tilde{c} by

$$(7.1) \quad \tilde{c}(\psi, a) := \int_{\mathbf{X}} c(x, a) \psi(dx), \quad \psi \in \mathcal{P}(\mathbf{X}).$$

It remains to construct the transition kernel \mathcal{K} . Working on the canonical sample space $\tilde{\Omega} = (\mathcal{P}(\mathbf{X}) \times \mathbf{A})^\infty$, we first define a stochastic kernel q on $\mathbf{X} \times \mathbf{Y}$ given $\mathcal{P}(\mathbf{X}) \times \mathbf{A}$ by

$$(7.2) \quad q(dx, dy \mid \psi, a) := \int_{\mathbf{X}} P(dx, dy \mid x', a) \psi(dx'), \quad \psi \in \mathcal{P}(\mathbf{X}),$$

and, decomposing q , we obtain

$$(7.3) \quad q(dx, dy \mid \psi, a) = Q(dy \mid \psi, a) \Psi(dx \mid \psi, a, y).$$

Equation (7.3) is the *filtering equation*. For fixed (ψ, a) , the map $y \mapsto \Psi$, as defined implicitly in (7.3), is a measurable mapping from \mathbf{Y} to $\mathcal{P}(\mathbf{X})$. Consequently, along

with the distribution Q on \mathbf{Y} , it induces a distribution \mathcal{K} on $\mathcal{B}(\mathcal{P}(\mathbf{X}))$, which is a measurable function of (ψ, a) or, in other words, a stochastic kernel on $\mathcal{P}(\mathbf{X})$ given $\mathcal{P}(\mathbf{X}) \times \mathbf{A}$. It follows that the model $(\mathcal{P}(\mathbf{X}), \mathbf{A}, \mathcal{K}, \tilde{c})$, with state process $\{\Psi_t\}_{t \in T}$, forms a completely observable controlled Markov process, with transition kernel given by

$$(7.4) \quad \mathcal{K}(B \mid \psi, a) := \int_{\mathbf{Y}} I\{\Psi(\cdot \mid \psi, a, y) \in B\} Q(dy \mid \psi, a), \quad B \in \mathcal{B}(\mathcal{P}(\mathbf{X})).$$

The distribution $\tilde{\mu}_0$ of Ψ_0 , corresponding to an initial distribution μ of the PO model, is taken to be

$$(7.5) \quad \tilde{\mu}_0(B) := \int_{\mathbf{Y}} \mu(B, dy), \quad B \in \mathcal{B}(\mathcal{P}(\mathbf{X})).$$

Given a history $\bar{h}_t = (\psi_0, y_0, \dots, a_{t-1}, y_t) \in \bar{\mathbf{H}}_t$ in the PO model, we can construct ψ_1, ψ_2, \dots in a recursive manner by starting from ψ_0 and, having obtained ψ_{t-1} , solving for Ψ in (7.3), with $(\psi, a, y) = (\psi_{t-1}, a_{t-1}, y_t)$, and letting $\psi_t = \Psi$. In this manner, we obtain a corresponding history $\tilde{h}_t = (\psi_0, a_0, \dots, a_{t-1}, \psi_t) \in \tilde{\mathbf{H}}_t$ for the CO model; we denote this correspondence by the map $g_t : \bar{\mathbf{H}}_t \rightarrow \tilde{\mathbf{H}}_t$. We can then assign to each admissible policy $\tilde{\pi} \in \tilde{\Pi}$ in the CO model a corresponding policy $\pi = g^*(\tilde{\pi})$ in the PO model, defined by

$$(7.6) \quad \pi_t(\cdot \mid \bar{h}_t) := \tilde{\pi}_t(\cdot \mid g_t(\bar{h}_t)), \quad \bar{h}_t \in \bar{\mathbf{H}}_t.$$

Clearly, every policy $\pi \in \Pi$ can also be regarded as a policy in $\tilde{\Pi}$; in other words, the map g^* is onto. If $\mathcal{P}_{\tilde{\mu}}^{\tilde{\pi}}$ is the probability measure induced by the policy $\tilde{\pi}$ and the initial distribution $\tilde{\mu}$ (corresponding to μ) on the canonical sample space $\tilde{\Omega}$, then, for each $C \in \mathcal{B}(\mathbf{X})$,

$$(7.7) \quad \mathcal{P}_{\mu}^{g^*(\tilde{\pi})}(X_t \in C \mid \bar{H}_t = \bar{h}_t) = \Psi_t(C), \quad \mathcal{P}_{\tilde{\mu}}^{\tilde{\pi}}\text{-a.s.}$$

Utilizing (7.1), (7.4), and (7.5), it can be verified that

$$(7.8) \quad E_{\mu}^{g^*(\tilde{\pi})}[c(X_t, A_t)] = E_{\tilde{\mu}}^{\tilde{\pi}}[\tilde{c}(\Psi_t, A_t)] \quad \forall t \in T,$$

thus establishing that the two models are indeed equivalent as claimed. It follows that the process Ψ_t summarizes all information, relevant for control purposes, and is called for this purpose a *sufficient statistic* (see [50], [161], [162]). We define the set of *separated policies* Π_S as those policies $\pi \in \Pi$ for which there a Markov policy $\tilde{\pi}$ on the equivalent CO problem such that $\pi = g^*(\tilde{\pi})$, as defined in (7.6). In other words, with $\tilde{\pi} = \{f_t\}_{t \in T} \in \tilde{\Pi}_M$, $f_t : \mathcal{P}(\mathbf{X}) \rightarrow \mathcal{P}(\mathbf{A})$ and for each initial distribution $\mu \in \mathcal{P}(\mathbf{S})$,

$$\pi_t(\cdot \mid \bar{h}_t) = f_t(\Psi_t)(\cdot), \quad \mathcal{P}_{\mu_0}^{\tilde{\pi}}\text{-a.s.}$$

Thus, the actions taken using a separated policy only depend on \bar{H}_t through the conditional distribution of X_t . In other words, the following *separation principle* holds: If an optimal policy exists in Π , one exists in Π_S . Hence, the process can be controlled optimally by first estimating the state via the conditional distribution and

choosing control actions based solely on the latter. These and other results, in various degrees of generality, were independently obtained by various authors, e.g., [3], [5], [89], [138], [151], [163], [174], [175], [199], [205].

Example 7.1. A partially observable version of the stochastic nonlinear system in Example 2.1 is described by the equations

$$\begin{aligned} X_{t+1} &= F(X_t, A_t, W_t), \\ Y_t &= G(X_t, A_{t-1}, V_t), \\ Y_0 &= G_0(X_0, V_0), \end{aligned}$$

where G and G_0 are Borel measurable, and the disturbance $\{V_t\}_{t \in T}$ is an i.i.d. sequence of random variables taking values in a Borel space \mathbf{V} , with a common distribution \mathcal{P}_V ; furthermore, it is assumed that X_0 , $\{W_t\}$, and $\{V_t\}$ are mutually independent.

7.2.2. Countable state space. We now specialize to the case where the state space $\mathbf{X} \times \mathbf{Y}$ is a finite or countably infinite set, the action space \mathbf{A} is a finite or compact set and with Assumption 7.1 in effect. Thus, $U(y) = \mathbf{A}$ for all $y \in \mathbf{Y}$, and the kernel of the process takes the form $P(x', y' \mid x, a)$. We also assume that the cost c and the kernel P are continuous with respect to $a \in \mathbf{A}$. The space $\mathcal{P}(\mathbf{X})$ is identified with the set Δ of probability vectors, i.e.,

$$(7.9) \quad \Delta := \left\{ \psi \in [0, 1]^{\mathbf{X}} : \sum_{x \in \mathbf{X}} \psi(x) = 1 \right\}$$

endowed with the topology given by the metric

$$d(\psi_1, \psi_2) := \sum_{x \in \mathbf{X}} |\psi_1(x) - \psi_2(x)| = \|\psi_1 - \psi_2\|_1,$$

where $\|\cdot\|_1$ stands for the standard ℓ_1 -norm on $\mathbb{R}^{\mathbf{X}}$.

In general, the recursive (filtering) equation (7.3) used to compute ψ_{t+1} , is obtained via a decomposition of measures technique; see [15, Chap. 10], [51, Chap. 8], [82, Chap. 4], [205]. This is particularly simple to accomplish (using the Bayes rule) when \mathbf{X} and \mathbf{Y} are countable or when the system is described by a linear system function and the disturbances are Gaussian; see [5], [14], [103], [174], [175]. For this purpose, we need the following definitions (compare with (7.2), (7.3)):

$$(7.10) \quad q(x, y \mid \psi, a) := \sum_{x' \in \mathbf{X}} P(x, y \mid x', a) \psi(x'),$$

$$(7.11) \quad V(y, \psi, a) := \sum_{x \in \mathbf{X}} q(x, y \mid \psi, a),$$

$$(7.12) \quad T(y, \psi, a)(\cdot) := \begin{cases} \frac{q(\cdot, y \mid \psi, a)}{V(y, \psi, a)}, & \text{if } V(y, \psi, a) \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the map $\psi \rightarrow T(y, \psi, a)$ maps Δ into itself. In the countable case, ψ_t can be computed by letting $\psi_t = T(y_t, \psi_{t-1}, a_{t-1})$. Here, $V(y, \psi, a)$ is interpreted as the (one-step ahead) conditional probability of the observation being y given an a priori distribution ψ for the core state, under decision a . Likewise, $T(y, \psi, a)$ is interpreted

as the a posteriori conditional probability distribution of the core state, given that decision a was made, observation y obtained, and an a priori distribution ψ . Also, the kernel in (7.4) takes the form

$$(7.13) \quad \mathcal{K}(B \mid \psi, a) := \sum_{y \in Y} V(y, \psi, a) I\{T(y, \psi, a) \in B\}, \quad B \in \mathcal{B}(\Delta),$$

while the cost \tilde{c} is computed by

$$(7.14) \quad \tilde{c}(\psi, a) := \sum_{x \in X} c(x, a) \psi(x).$$

Remark 7.3. It is common to specify, instead of the kernel P , a transition kernel \bar{P} on X given $X \times A$, and an *observation kernel* \bar{Q} on Y given $X \times A$ [14], [63], [82], [128], [170]. Note that this is only a special case of our presentation, which happens when the kernel P admits the decomposition

$$P(x, y \mid x', a) = \bar{Q}(y \mid x, a) \bar{P}(x \mid x', a).$$

In this case, we can express (7.10)–(7.12) in a convenient vector form by viewing ψ as an element of \mathbb{R}^X and defining the transition matrix $[\bar{P}(a)]_{x, x'} := \bar{P}(x \mid x', a)$ and the observation matrix $Q_y(a) := \text{diag}\{Q(y \mid x, a) : x \in X\}$. Then, with \bar{q} denoting the vector in \mathbb{R}^X defined by $\bar{q}_x(y \mid \psi, a) := q(x, y \mid \psi, a)$ and $\mathbf{1}' = (1, \dots, 1)$, we have

$$(7.10') \quad \bar{q}(y \mid \psi, a) = \bar{Q}_y(a) \bar{P}(a) \psi,$$

$$(7.11') \quad V(y, \psi, a) = \mathbf{1}' \bar{Q}_y(a) \bar{P}(a) \psi$$

(analogously for (7.12)).

Note that a *nonrandomized* separated admissible policy can be viewed as a sequence of maps $\pi_t : \Delta \rightarrow A$. Then an equivalent, *completely observable*, discounted cost problem (DC') can be formulated as finding a separated admissible policy that minimizes

$$J_\beta(\psi, \pi) := E_{\psi_0}^\pi \left[\sum_{t=0}^{\infty} \beta^t \tilde{c}(\Psi_t, A_t) \right].$$

The average cost problem (AC') is analogously defined.

Note that the one-stage cost function $\tilde{c}(\psi, a)$ is linear in $\psi \in \Delta$. It is easy to show that the expectation operator corresponding to the kernel \mathcal{K} preserves concavity (convexity) [6], [50]. The following results complement those in Theorem 2.1.

THEOREM 7.1. *For a (DC') decision problem, $J_\beta^*(\cdot)$ is a concave function, for all $0 < \beta < 1$. The DCOE is given by*

$$(7.15) \quad J_\beta^*(\psi) = \min_{a \in A} \left\{ \tilde{c}(\psi, a) + \beta \sum_{y \in Y} V(y, \psi, a) J_\beta^*(T(y, \psi, a)) \right\},$$

and any (nonrandomized) separated stationary policy that attains the minimum above is optimal.

Remark 7.4. The optimality equation (7.15) is obtained from the general theory of CMP [15], [82]. For other results, see [5]–[7], [14], [50], [128], [161], [169], [170], [171]. Also, for a survey of relevant computational methods, see [119].

In this context, a pair (ρ, h) is said to be a solution to the ACOE if, for all $\psi \in \Delta$,

$$(7.16) \quad \rho + h(\psi) = \min_{a \in \mathbf{A}} \left\{ \tilde{c}(\psi, a) + \sum_{y \in \mathbf{Y}} V(y, \psi, a) h(T(y, \psi, a)) \right\}.$$

7.3. The vanishing discount approach. As shown in §5, for a countable state space CMP, boundedness conditions on the differential discounted value function were sufficient for solutions to the corresponding ACOE to exist. We consider here the following hypothesis.

Assumption 7.2. There exists a sequence $\beta_n \uparrow 1$, such that h_{β_n} is bounded.

Despite the fact that the model $(\Delta, \mathbf{A}, \mathcal{K}, \tilde{c})$ has a general Borel state space, it has two special features that simplify the analysis via the vanishing discount method. The first of these features is the concavity of the discounted value function, while the second is the fact that the kernel $\mathcal{K}(\cdot \mid \psi, a)$ vanishes on the complement of a countable set (for fixed ψ and a), and thus the integrals with respect to \mathcal{K} reduce to infinite sums.

For the finite state and action space case, the concavity of the discounted value function has been exploited by Platzman [136] and by Ohnishi, Mine, and Kawai [132]. These authors utilize the fact that a collection of concave functions, defined on some relatively open convex set C , which are finite and pointwise bounded, is uniformly bounded and equi-Lipschitzian relative to any closed subset of C [143, Thm. 10.6]. Thus, under Assumption 7.2, the finite dimensionality of Δ and the concavity of $h_\beta(\cdot)$ are used in [132], [136] to obtain a bounded solution (ρ^*, h) to the ACOE, via the vanishing discount approach. In particular, they partition Δ into its interior, its vertices, and its edges, i.e.,

$$\Delta = \bigcup_{j \in \mathcal{J}} \Delta_j.$$

Note that $|\mathcal{J}| = 2^{|\mathbf{X}|+1} - 1$ and that each set Δ_j is a relatively open convex set. Given a sequence $\beta_n \uparrow 1$, then the concavity of $h_\beta(\cdot)$ and Assumption 7.2 are used to obtain subsequences $\beta_n(j)$ such that $\{h_{\beta_n(j)}(\cdot)\}$ converges on Δ_j . Platzman [136] defines a metric on Δ that accomplishes this partition. Let

$$\begin{aligned} \mathcal{I}(\psi) &:= \{i \in \mathbf{X} : \psi(i) > 0\}, \quad \psi \in \Delta, \\ d(\psi_1, \psi_2) &:= 1 - \min \left\{ \frac{\psi_1(i)}{\psi_2(i)} : i \in \mathcal{I}(\psi_2) \right\}, \quad \psi_1, \psi_2 \in \Delta, \\ D(\psi_1, \psi_2) &:= \max \{d(\psi_1, \psi_2); d(\psi_2, \psi_1)\}. \end{aligned}$$

In [135, pp. 88–89], Platzman shows that $D(\cdot, \cdot)$ is a metric that leaves Δ disconnected and with components identical to the elements of the partition $\{\Delta_j\}_{j \in \mathcal{J}}$. The following is shown in [136, Lemma A.1].

LEMMA 7.1. *Let $f : \Delta \rightarrow \mathbb{R}$ be concave and bounded below; then*

$$|f(\psi_1) - f(\psi_2)| \leq \text{span}(f) D(\psi_1, \psi_2).$$

Hence, under Assumption 7.2, $\{h_\beta(\cdot)\}_{\beta \in (0,1)}$ is an equi-Lipschitzian family, with common Lipschitz constant given by the (smallest) uniform bound, and the Arzela–Ascoli theorem can be used as in [148] to obtain a bounded solution to the ACOE.

If the state space is infinite, the above method does not work, simply because the partition induced by the Platzman metric results in a nonseparable space. In this situation, the particular structure of the kernel has been employed in [63] to develop a theoretical framework based on the notion of *invariant* subsets (subprocesses) of a CMP, and sufficient conditions are given for the existence of solutions to the ACOE, in the case of a finite action space. The key point is to note that, if we let $B(\psi, a) := \{T(y, \psi, a) : y \in \mathbf{Y}\}$, which is a countable set since \mathbf{Y} is countable, then $\mathcal{K}(B(\psi, a) | \psi, a) = 1$. Thus, at any time $t \in \mathbb{N}_0$, the set of possible *next states* for Ψ_t is the set $\bigcup_{a \in \mathbf{A}} B(\Psi_t, a)$, which is countable, provided that \mathbf{A} is finite. This special structure has also been identified by other authors, e.g., [5, p. 187], [136, p. 369], [170, pp. 19–20].

We briefly summarize the work in [63]. The notions of *descendents*, *ancestors*, and *relatives* of a point $\psi \in \Delta$ are first introduced. The descendents of ψ are defined as the smallest subset of Δ containing ψ that is invariant under the action of the maps in the collection $\{T(y, \cdot, a) : y \in \mathbf{Y}, a \in \mathbf{A}\}$, while the ancestors of ψ are defined as all the points in Δ that reach ψ under the application a finite sequence of these maps. Finally, the relatives of a point ψ , denoted by $\mathcal{R}_\psi^{(1)}$, is the set formed by the union of its descendents and ancestors. Note that the definition of the descendents is an extension, to the present context, of Doob's concept of *consequent sets* [45, p. 206]. Subsequently, the *genealogical tree* GT_ψ of ψ is defined by

$$GT_\psi := \bigcup_{n \in \mathbb{N}} \mathcal{R}_\psi^{(n)},$$

where the sets $\mathcal{R}_\psi^{(n)}$ are defined recursively as

$$\mathcal{R}_\psi^{(n+1)} := \bigcup_{s \in \mathcal{R}_\psi^{(n)}} \mathcal{R}_s^{(1)}, \quad n \in \mathbb{N}.$$

The descendents of a point form a countable set, but the ancestors can, in general, be uncountably many. To guarantee that the relatives and hence the genealogical tree of a point is a countable set, the following condition is introduced.

Assumption 7.3. For all $y \in \mathbf{Y}$, $a \in \mathbf{A}$, and $\psi \in \Delta$, $T^{-1}(y, \psi, a)$ is a countable set.

Introduce the relation $\psi \sim \psi'$ if $GT_\psi = GT_{\psi'}$. It follows that “ \sim ” defines an *equivalence relation* on Δ resulting in a partition of Δ into equivalence classes that are precisely the sets GT_ψ . Under Assumptions 7.2 and 7.3, the standard diagonalization argument can be employed on each equivalence class GT_ψ to construct a pair (ρ^*, h_{GT_ψ}) that solves the ACOE on GT_ψ (the boundedness hypothesis (Assumption 7.2) can be weakened by letting the constant M depend on the equivalence class). Then, by defining $h(\psi) := h_{GT_\psi}(\psi)$ for all $\psi \in \Delta$, (ρ^*, h) clearly solves the ACOE on Δ . One peculiarity of this approach is that the resulting function h is not guaranteed to be measurable. This is not a major problem though, since an important consequence of the particular structure (with finite action space) is that the “measurability of various objects is of no essential concern” for the equivalent problem [15, p. xi]. The approach in [63] fails when the action space \mathbf{A} is not finite.

Since the vanishing discount method relies heavily on the boundedness of the differential discounted value function, the problem of finding sufficient conditions on the cost and the kernel of the process for this to hold becomes important. Platzman [136] has given (reachability and detectability) conditions for Assumption 7.2 to hold;

however, these conditions are difficult to verify. On the other hand, many models of interest possess special properties, which allow the verification of Assumption 7.2 very easily. We examine some of these properties next.

Suppose that a partial order “ \prec_{Δ} ” has been defined on Δ and let “ \prec_A ” denote a linear order on A ; we assume that A is finite. We also identify X with \mathbb{N}_0 and endow it with its natural ordering.

DEFINITION 7.1. Consider $((\Delta, \prec_{\Delta}), (A, \prec_A), \mathcal{K}, \tilde{c})$ and let $\psi_1, \psi_2 \in \Delta$. We state the following:

(i) The value functions are *monotone* if

$$\psi_1 \prec_{\Delta} \psi_2 \implies J_{\beta}^*(\psi_1) \leq J_{\beta}^*(\psi_2) \quad \text{for all } 0 < \beta < 1;$$

(ii) A (nonrandomized) stationary separated policy π is *monotone* if

$$\psi_1 \prec_{\Delta} \psi_2 \implies \pi(\psi_1) \prec_A \pi(\psi_2).$$

Two frequently used partial orders on Δ are the *stochastic dominance* \prec_{st} and the *monotone likelihood ratio* \prec_{lr} , defined below.

DEFINITION 7.2. Let $\psi_1, \psi_2 \in \Delta$; we state the following:

- (i) $\psi_1 \prec_{st} \psi_2$ if $\sum_{i \geq q} \psi_1(i) \leq \sum_{i \geq q} \psi_2(i)$, for all $q \in X$;
- (ii) $\psi_1 \prec_{lr} \psi_2$ if $\psi_1(j)\psi_2(i) \leq \psi_1(i)\psi_2(j)$, for all $i, j \in X$ such that $i \leq j$.

Let e^j denote the element of Δ with the j th component equal to 1, $j \in X$; thus, e.g., $e^0 = (1, 0, 0, \dots)$. The following is easily shown.

LEMMA 7.2. If $\psi_1, \psi_2 \in \Delta$ and $\psi_1 \prec_{lr} \psi_2$, then $\psi_1 \prec_{st} \psi_2$. Also, for all $\psi \in \Delta$, $e^0 \prec_{lr} \psi$.

DEFINITION 7.3. An action $a_j \in A$ is called a reset action if, for some $j \in X$, $T(y, \psi, a_j) = e^j$, for all $y \in Y$ and $\psi \in \Delta$.

A reset action a_j corresponds to the core state of the system being j , with probability one, at the next time epoch after action a_j has been taken. This type of action arises naturally in manufacturing systems subject to inspection, maintenance, and replacement. The following results derive from the work of Sondik [170]; see also [63].

LEMMA 7.3. If there exists a reset action $a_j \in A$, then

$$J_{\beta}^*(\psi) - J_{\beta}^*(e^j) \leq \tilde{c}(\psi, a_j) \quad \forall \psi \in \Delta.$$

If X is finite and for each $j \in X$ there is a corresponding reset action, then for each $\beta \in (0, 1)$ there exists $J \in X$ such that

$$0 \leq J_{\beta}^*(\psi) - J_{\beta}^*(e^J) \leq M \quad \forall \psi \in \Delta,$$

where $M := \max\{c(i, a) \mid i \in X, a \in A\}$.

Remark 7.5. Note that, if $J_{\beta}^*(\cdot)$ is monotone with respect to \prec_{lr} and if there is an action $a_0 \in A$ that resets the state to e^0 , then $0 \leq J_{\beta}^*(\psi) - J_{\beta}^*(e^0) \leq \tilde{c}(\psi, a_0)$ uniformly in $\beta \in (0, 1)$. Furthermore, note that when X is finite, a constant $M > 0$ exists such that $\tilde{c}(\psi, a_0) \leq M$, for all $\psi \in \Delta$, and thus Assumption 7.2 holds.

Models with a *replacement* action that resets the system to an “as new” state e^0 have been considered in [2], [118], [131], [132], [149], [188], [189], [191]–[195]. Related problems are those considered in [66], where a reset action to a most desirable state is available, and in [90], where (maintenance) reset actions a_j are available for all $j \neq 0$, with X a finite set.

7.4. The convex analytic method. We will now briefly describe Borkar's convex analytic approach. The action set \mathbf{A} is assumed to be any compact metric space. We also assume that c and P are continuous in a . We will consider the pathwise average cost. This cannot, in general, be written as an equivalent cost in terms of $\{\Psi_t\}$, but it is natural to propose that

$$(7.17) \quad \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{c}(\Psi_t, A_t)$$

as a substitute. Any $\mu \in \mathcal{P}(\Delta \times \mathbf{A})$ can be decomposed as

$$(7.18) \quad \mu(d\psi, da) = \bar{\mu}(d\psi) \Phi(\psi)(da),$$

where $\bar{\mu}$ is the marginal of μ on Δ and Φ is the regular conditional law defined $\bar{\mu}$ almost surely. We always work with one arbitrary representative of this equivalence class. Define $\Gamma \subset \mathcal{P}(\Delta \times \mathbf{A})$ by

$$(7.19) \quad \begin{aligned} \Gamma &= \left\{ \mu \in \mathcal{P}(\Delta \times \mathbf{A}) \mid \begin{array}{l} \text{For } \bar{\mu}, \Phi \text{ as in (7.18), } \bar{\mu} \text{ is invariant under } \\ \text{the stationary randomized policy } \Phi \end{array} \right\} \\ &= \left\{ \mu \in \mathcal{P}(\Delta \times \mathbf{A}) \mid \iint f(\psi) \mathcal{K}(d\psi \mid \psi', a) \Phi(\psi')(da) \bar{\mu}(d\psi') \right. \\ &\quad \left. = \int f d\bar{\mu} \text{ for all } f \in C_b(\Delta) \right\}. \end{aligned}$$

From (7.19) we can easily check that Γ is closed. Note that the set of invariant probability measures for the process $\{\Psi_t\}$ controlled by a stationary randomized policy Φ , when nonempty, need not be a singleton. In general, it will form a closed convex set in $\mathcal{P}(\Delta)$, the extreme points of which correspond to ergodic measures. That is, the above process with one of these extreme measures (say, μ) as the initial condition will be ergodic. Then (7.17) will almost surely equal $\int \tilde{c} d\mu$. In view of the ergodic decomposition of a stationary Markov process, this will also be the case for other invariant measures (which will be a convex combination of the ergodic ones). Define

$$\rho^* = \inf_{\mu \in \Gamma} \int \tilde{c} d\mu.$$

We assume that $\rho^* < \infty$. We consider two alternative conditions under which the above infimum will be a minimum.

Assumption 7.4 (near-monotone case). c satisfies $\lim_{i \rightarrow \infty} \inf_a c(i, a) = \infty$.

Assumption 7.5 (stable case). Assumption 5.19' (ii) holds.

Observe that the “near-monotonicity” condition here is more restrictive than the one used in §5. We now state the following result; the proof is analogous to that of Theorem 5.10.

LEMMA 7.4. *Under either Assumption 7.4 or Assumption 7.5, the map $\mu \mapsto \int \tilde{c} d\mu$ attains its minimum on Γ .*

Define the $\mathcal{P}(\Delta \times \mathbf{A})$ -valued process $\{\eta_t\}$ by

$$\int f d\eta_t = \frac{1}{t} \sum_{m=0}^{t-1} f(\Psi_m, A_m), \quad t \geq 1, \quad f \in C_b(\Delta \times \mathbf{A}),$$

where $\{\Psi_t\}$ is governed by some policy. Again, we can prove the following analogue of Lemma 5.1.

LEMMA 7.5. *With probability 1, any limit point of $\{\eta_t\}$ in $\mathcal{P}(\Delta \times \mathbf{A})$ lies in Γ .*

Consider the near-monotone case. Suppose that, for a given sample path, a subsequence of $\{\eta_t\}$ has no limit point in $\mathcal{P}(\Delta \times \mathbf{A})$. Arguments similar to those in the proof of Theorem 5.1 can be used to show that the cost must go to $+\infty$ along this subsequence. In view of Lemma 7.5, this leads to

$$(7.20) \quad \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \tilde{c}(\Psi_t, A_t) \geq \rho^* \quad \text{a.s.}$$

Along with Lemma 7.4, this would seem to lead to the existence of an optimal stationary randomized policy. There is, however, one catch. It is not a priori clear that any initial law for $\{\Psi_t\}$ would be in the domain of attraction of the element(s) of Γ that minimize the cost (or, for that matter, whether this domain of attraction can be reached in a finite random time from any initial law under some policy). Similar “reachability” problems surface when we try to extend the dynamic programming equations. These are circumvented under somewhat stringent conditions in [136], as we have already discussed.

Finally, we can prove the convexity of Γ . Again, it is unclear how (and whether) we can characterize the extreme points of Γ as those corresponding to stationary (nonrandomized) policies. As the ACOE is not available in this approach, the existence of an optimal stationary policy remains an open issue in general. In the stable case, it is not clear if (7.20) holds, and thus this case remains open to investigation. To sum up, the convex analytic approach to POCMP needs to be further studied.

8. Multiobjective and constrained models. An important success of the convex analytic approach discussed in §5 is in the domain of multiobjective problems, in which there is more than one cost (objective) function. We will first consider a multiobjective CMP with average cost criterion recast as a CMP with several constraints. CMP with one or multiple constraints have been studied in [1], [16], [26], [40], [41], [44], [92], [93], [97], [120], [129], [145], [146], [166]. Our presentation follows [25], [26].

We consider the case when $\mathbf{S} = \{0, 1, 2, \dots\}$; \mathbf{A} , the action space, is a prescribed compact metric space; and $P(j | i, a)$ is continuous in a for fixed i, j . Also, $U(i) = \mathbf{A}$ for all $i \in \mathbf{S}$. In the constrained CMP problem, we have, in addition to the cost function $c \in C_b(\mathbf{S} \times \mathbf{A})$, m additional “costs” $c_i \in C_b(\mathbf{S} \times \mathbf{A})$, $1 \leq i \leq m$ and are required to satisfy

$$(8.1) \quad a_i \leq \int c_i d\hat{\eta}(f) \leq b_i, \quad 1 \leq i \leq m$$

for prescribed numbers $b_i > a_i$, $f \in \Pi_{SD}$, and $\hat{\eta}(f) \in \mathcal{P}(\mathbf{S} \times \mathbf{A})$ is as in §5. (We are assuming all costs are bounded for simplicity. Also, we are confining our attention to Π_{SSR} ; this suffices under reasonable hypotheses, as we saw in §5.) We will assume Assumption 5.20 in §5.

Recall that $I_R = \{\hat{\eta}(f) : f \in \Pi_{SSR}\}$. Let \tilde{I}_R be the subset of I_R , where the constraints (8.1) are satisfied. Then \tilde{I}_R is closed and convex. We assume also that it is compact (this will be true under Assumption 5.20 in §5). Under this assumption, we can show, as in §5, that there exists an $f^* \in \Pi_{SR}$ that is optimal for this problem. We will now proceed to show that f^* requires randomization in at most m states.

Let $g \in C_b(\mathbf{S} \times \mathbf{A})$. For some $a \in \mathbb{R}$, let $H = I_R \cap \{\psi : \int g d\psi \leq a\}$, assumed to be nonempty. Clearly, H is closed and convex. Let $\hat{\eta}(f)$ be an extreme point of H .

Suppose that it is not an extreme point of I_R itself. Then there exist distinct measures $\hat{\eta}(f_{11})$, $\hat{\eta}(f_{12})$ such that at least one of them (say $\hat{\eta}(f_{11})$) is not in H , and $\hat{\eta}(f)$ is a convex combination of the two. Suppose that $\hat{\eta}(f_{21}) \in I_R \setminus H$, $\hat{\eta}(f_{22})$ is another such pair. Then it can be shown that $\hat{\eta}(f_{ij})$, $1 \leq i, j \leq 2$ are collinear (I_R , \tilde{I}_R , H , and so on are viewed as subsets of $\mathfrak{M}(\mathbf{S} \times \mathbf{A})$, the Banach space of finite signed measures on $\mathbf{S} \times \mathbf{A}$). Therefore, all pairs of points in I_R satisfying (a) at least one of them is not in H , and (b) $\hat{\eta}(f)$ is a convex combination thereof, lie on a single straight line in $\mathfrak{M}(\mathbf{S} \times \mathbf{A})$. Let Z denote the intersection of this line with I_R . Under our hypotheses on I_R , Z is a closed finite line segment. Let $\eta(f_1)$, $\eta(f_2)$ denote its endpoints. Then it can be shown that $\eta(f_i)$, $i = 1, 2$ are extreme points of I_R . By Lemma 5.2, $f_i \in \Pi_{SSD}$; also, f_1 and f_2 are distinct since $\hat{\eta}(f)$ is not an extreme point of I_R . Therefore, there exists an $a' \in (0, 1)$ such that

$$\hat{\eta}(f) = a'\hat{\eta}(f_1) + (1 - a')\hat{\eta}(f_2).$$

Arguing as in the proof of Lemma 5.2, it is clear that for each $i \in \mathbf{S}$ we may take $f(i)$ to be a convex combination of $f_1(i)$ and $f_2(i)$. Let $\tilde{f} \in \Pi_{SD}$ be such that, for each $i \in \mathbf{S}$, $\tilde{f}(i) =$ either $f_1(i)$ or $f_2(i)$. Then, under our hypotheses (Assumption 5.20 of §5), we can show that $\hat{\eta}(\tilde{f}) \in Z$. Now consider Z as a union of two closed line segments Z_1 and Z_2 , Z_1 being the line segment between $\hat{\eta}(f_1)$ and $\hat{\eta}(f)$, and Z_2 that between $\hat{\eta}(f_2)$ and $\hat{\eta}(f)$. Let $\{f'_n\}$ be a sequence in Π_{SD} , defined as follows: $f'_0 = f_1$, and

$$f'_n(i) = \begin{cases} f_2(i), & i \leq n, \\ f_1(i), & i > n. \end{cases}$$

Then, by the above considerations, $\hat{\eta}(f'_n) \in Z$. Since $f'_n \rightarrow f_2$ as $n \rightarrow \infty$, we conclude that $\hat{\eta}(f'_n) \rightarrow \hat{\eta}(f_2)$ (the map $f \mapsto \hat{\eta}(f)$ is continuous under Assumption 5.19). Thus, the sequence $\hat{\eta}(f'_n)$, $n \geq 0$ starts in Z_1 and eventually moves into Z_2 . Let n denote the first time this happens. Then either $\hat{\eta}(f'_n) = \hat{\eta}(f)$ or $\hat{\eta}(f)$ is a convex combination of $\hat{\eta}(f'_n)$ and $\hat{\eta}(f'_{n-1})$. Since $f'_n(i) = f'_{n-1}(i)$ for $i \neq n$, the arguments employed in Lemma 5.2 show that we may take $f(i) =$ the Dirac measure at $f'_n(i)$ for $i \neq n$ and $f(n) =$ a suitable convex combination of Dirac measures at $f_1(n)$ and $f_2(n)$. We have established the following result.

THEOREM 8.1. *Each extreme point of H corresponds to an $\hat{\eta}(f)$ such that $f \in \Pi_{SR}$ satisfies the following claim: For all but at most one i , $f(i)$ is a Dirac measure at some point of \mathbf{A} . For the single remaining i , if any, $f(i)$ is a convex combination of two such Dirac measures.*

A variant of the above theorem leads to the following result [27].

THEOREM 8.2. *The minimum of $\nu \mapsto \int c d\nu$ on \tilde{I}_R , is attained at an $\hat{\eta}(f) \in \tilde{I}_R$, where f is either deterministic or satisfies the following claim: There are states $i_1, \dots, i_k \in \mathbf{S}$ and positive integers $n_1, \dots, n_k > 1$ such that f requires randomization among n_j values at state i_j , $1 \leq j \leq k$; requires no randomization for the remaining states; and $\sum_{i=1}^k n_i \leq m$.*

Once this existence result is available, necessary conditions for optimality can be obtained from the standard Lagrange multiplier theory.

THEOREM 8.3. *There exist $\lambda_i, \beta_i \geq 0$, $1 \leq i \leq k$ such that $\hat{\eta}(f)$, as in Theorem 8.2, minimizes*

$$\eta \mapsto F(\eta, \{\lambda_i\}, \{\beta_i\}) := \int c d\eta - \sum_{i=1}^k \lambda_i (b_i - \int c_i d\eta) - \sum_{i=1}^k \beta_i (\int c_i d\eta - a_i)$$

on I_R . Furthermore, if \tilde{I}_R has nonempty interior, the following saddle-point property holds: For all $\bar{\lambda}_i, \bar{\beta}_i \geq 0$, $1 \leq i \leq k$, $\eta \in I_R$,

$$F(\hat{\eta}(f), \{\bar{\lambda}_i\}, \{\bar{\beta}_i\}) \leq F(\hat{\eta}(f), \{\lambda_i\}, \{\beta_i\}) \leq F(\eta, \{\lambda_i\}, \{\beta_i\}).$$

Remark 8.1. The result in Theorem 8.1 cannot be improved in general. Indeed, in [26] there is a counterexample to show the nonexistence of an optimal $f \in \Pi_{SD}$ for the CMP with one constraint.

Remark 8.2. We have discussed the stable case only. Analogous results can be obtained for the near-monotone case (conditions similar to Assumption 5.18). For details, we refer to [25].

Remark 8.3. When the action set \mathbf{A} is countable, analogous results are obtained in [1].

We next consider another multiobjective CMP with AC criterion. We have m cost functions $c_i \in C_b(\mathbf{S} \times \mathbf{A})$, $1 \leq i \leq m$. All cost functions are of equal importance, and, as a result, the optimality problem cannot be recast as a constrained one. Therefore, we directly deal with the optimality problem with a vector cost criterion. This has been studied in [48], [75].

Let I_R be compact. Consider the vector cost criterion

$$\left(\int c_1 d\hat{\eta}(f), \dots, \int c_m d\hat{\eta}(f) \right), \quad \hat{\eta}(f) \in I_R.$$

In general, there need not exist an $f \in \Pi_{SSR}$ that minimizes all of $\int c_i d\hat{\eta}(f)$ over I_R . This motivates the concept of Pareto optimality. An $f \in \Pi_{SSR}$ is said to be *Pareto optimal* if there does not exist any $\bar{f} \in \Pi_{SSR}$ for which $\int c_i d\hat{\eta}(\bar{f}) \leq \int c_i d\hat{\eta}(f)$, $1 \leq i \leq m$, with inequality being strict for at least one i . Pareto optimality is clearly the minimal requirement for any reasonable notion of an optimal solution for the multiobjective problem with no priority among objectives. The Pareto optimal solutions can be characterized as follows.

THEOREM 8.4. *Any $f \in \Pi_{SSR}$ that minimizes $\sum_{i=1}^m \lambda_i \int c_i d\hat{\eta}(f)$ for some $\lambda_i > 0$, $1 \leq i \leq m$ is Pareto optimal. Conversely, any Pareto optimal $\bar{f} \in \Pi_{SSR}$ minimizes the above functional for some choice of $\lambda_i \geq 0$, $1 \leq i \leq m$.*

Remark 8.4. Note that the converse is only partial, since we have $\lambda_i \geq 0$ instead of $\lambda_i > 0$. It becomes exact if \mathbf{S} and \mathbf{A} are finite.

We often reduce a vector cost criterion as above to a scalar one by introducing a “utility function.” One such case is that of finding the “shadow minimum” for the problem of minimizing the vector cost $\nu \mapsto [\int c_1 d\nu, \dots, \int c_m d\nu] \in \mathbb{R}^m$ on I_R . Letting L denote the range of this map, L can be shown to be closed and convex. Suppose that $y_i^* = \min\{\int c_i d\nu : \nu \in I_R\}$, $1 \leq i \leq m$. Let $y^* = (y_1^*, \dots, y_m^*)$. The point y^* is called the ideal (or utopian) point. The point $x^* \in L$ that is closest to y^* is called the *shadow minimum*. This point is unique and is characterized by

$$\langle y^* - x^*, z - x^* \rangle \leq 0, \quad z \in \mathbf{S}.$$

For finite \mathbf{S} and \mathbf{A} , a combined linear-quadratic program can find x^* explicitly [75]. The point x^* is easily seen to be Pareto optimal.

9. Conclusions. We hope this paper has provided a useful presentation of the problems and techniques in average cost control of Markov processes. As is amply

clear, there is not a globally applicable approach. Instead, we expect to build a library of special tricks, a collection of simple verifiable sufficient conditions under which the problem is accessible, possibly with different techniques. Going one step further, there are the more difficult, partially observable, and multiobjective problems. Though these have seen some significant results of late, there remains much more that eludes satisfactory analysis. A similar comment applies to computational aspects and adaptive control, two topics we have not touched upon here. For computational aspects, we refer to [81], [87], [137], [180] and, for adaptive control, [26], [82], [102]. Also, we have not dealt with the vast literature on *sensitive* optimality [137], [182], nor with some other criteria, such as overtaking [111], variance sensitive [198], and weighted cost [60], [65], [99]. Finally, the discrete-time models have interesting applications to continuous-time problems, for which we refer to [14, §6.7], [109], [159], [206].

Appendix. Multifunctions and measurable selectors. Let V and W denote nonempty Borel spaces and let 2^W denote the collection of all *nonempty* subsets of W . A *multifunction* (or set-valued function) Φ from V to W is a map $\Phi : V \rightarrow 2^W$. The subset $\text{Dom}(\Phi) := \{v \in V : \Phi(v) \neq \emptyset\}$ is called the *domain* of Φ . When $\text{Dom}(\Phi) = V$, we say that the map Φ is *strict*. In what follows, we assume that Φ is a strict multifunction. If, for each $v \in V$, $\Phi(v)$ is a compact (closed, measurable) subset of W , then Φ is said to be *compact (closed, measurable)-valued*. A *selector* (or selection) of Φ is a function $\varphi : V \rightarrow W$ such that $\varphi(v) \in \Phi(v)$, for all $v \in \text{Dom}(\Phi)$. The set of (Borel) measurable selectors of Φ will be denoted by $S(\Phi)$. The *graph* of Φ , denoted by $\text{Graph}(\Phi)$, is defined as

$$\text{Graph}(\Phi) := \{(v, w) : v \in V, w \in \Phi(v)\}.$$

For a set $W \in 2^W$, we define

$$\Phi^{-1}[W] := \{v \in V : \Phi(v) \cap W \neq \emptyset\},$$

and we say that Φ is (Borel) *measurable* if $\Phi^{-1}[B] \in \mathcal{B}(V)$, for each *closed* subset B of W . If Φ is *closed-valued*, then measurability of Φ implies that $\text{Graph}(\Phi) \in \mathcal{B}(V \times W)$; furthermore, if Φ is *compact-valued*, then the converse also holds [88], [184, Thm. 4.2]. The multifunction Φ is called *upper semicontinuous* if, for every $v \in V$ and every open set $G \supset \Phi(v)$, there exists a neighborhood N of v such that $\Phi(v') \subset G$, for all $v' \in N$; it is called *lower semicontinuous* if, for every $v \in V$ and every open set G such that $G \cap \Phi(v) \neq \emptyset$, $\Phi^{-1}(G)$ contains an open neighborhood of v . Also, Φ is said to be *continuous* if it is both upper and lower semicontinuous.

The following result, in different variations, has been shown by several authors [15, §7.5], [47, Lemma 6, p. 38], [51, Chap. 2], [88], [154] and also summarized in [82], [184, Thm. 9.1].

THEOREM A.1. *Let Φ be a compact-valued, measurable, strict multifunction from V to W . Let $f : \text{Graph}(\Phi) \rightarrow \mathbb{R}$ be a measurable function, such that, for each $v \in V$, $f(v, \cdot)$ is lower semicontinuous on $\Phi(v)$. Then there exists a measurable selector $\varphi^* \in S(\Phi)$ such that*

$$f(v, \varphi^*(v)) = \min_{w \in \Phi(v)} \{f(v, w)\} \quad \forall v \in V.$$

Let $f^ : V \rightarrow \mathbb{R}$, defined by $f^*(v) := f(v, \varphi^*(v))$. If Φ is upper semicontinuous and f is bounded below, then $f^* \in \mathcal{L}(V)$. Also, if Φ is continuous and $f \in C_b(V \times W)$, then $f^* \in C_b(V)$.*

A Tauberian theorem. The following Tauberian theorem plays a very important role in the analysis of the average cost criterion. For its proof, which is very difficult to locate in the literature in this particular format, we refer to [176].

THEOREM A.2. *Let $\{a_n\}$ be a sequence of nonnegative numbers and $\beta \in (0, 1)$. Then*

$$\begin{aligned} \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m &\leq \liminf_{\beta \uparrow 1} (1 - \beta) \sum_{n=0}^{\infty} \beta^n a_n \\ &\leq \limsup_{\beta \uparrow 1} (1 - \beta) \sum_{n=0}^{\infty} \beta^n a_n \leq \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m. \end{aligned}$$

Acknowledgments. The authors would like to thank the anonymous referee and Associate Editor Steven E. Shreve for their constructive criticism and comments, which helped to improve this paper. Our thanks also go to Prof. Linn I. Sennott for pointing out an error in the statement of Theorem 5.1, in an earlier draft of this manuscript. We were blessed by having an excellent typist, Joan Van Cleave, who rose above mere patience when faced with numerous revisions of our “final draft.” Finally, E. Fernández-Gaucherand wishes to thank Prof. O. Hernández-Lerma of CINVESTAV-IPN, México for useful discussions.

REFERENCES

- [1] E. ALTMAN AND A. SHWARTZ, *Markov decision problems and state-action frequencies*, SIAM J. Control Optim., 29 (1991), pp. 786–809.
- [2] V. A. ANDRIYANOV, I. A. KOGAN AND G. A. UMNNOV, *Optimal control of a partially observable discrete Markov process*, Automat. Remote Control, 4 (1980), pp. 555–561.
- [3] M. AOKI, *Optimal control of partially observable Markovian systems*, J. Franklin Inst., 280 (1965), pp. 367–386.
- [4] K. J. ARROW, T. HARRIS, AND J. MARSHAK, *Optimal inventory policy*, Econometrica, 19 (1951), pp. 250–272.
- [5] K. J. ÅSTRÖM, *Optimal control of Markov processes with incomplete state information*, J. Math. Anal. Appl., 10 (1965), pp. 174–205.
- [6] ———, *Optimal control of Markov processes with incomplete state information, II. The convexity of the loss function*, J. Math. Anal. Appl., 26 (1969), pp. 403–406.
- [7] ———, *Stochastic control problems*, in Mathematical Control Theory, W. A. Coppel, ed., Lecture Notes in Mathematics, Vol. 680, Springer-Verlag, Berlin, 1978, pp. 1–69.
- [8] J. A. BATHER, *Optimal decision procedures for finite Markov chains*, I: *Examples*, Adv. Appl. Probab., 5 (1973), pp. 328–339; II: *Communicating systems*, Adv. Appl. Probab., 5 (1973), pp. 521–540; III: *General convex systems*, Adv. Appl. Probab., 5 (1973), pp. 541–553.
- [9] R. BELLMAN, *A Markovian decision problem*, J. Math. Mech., 6 (1957), pp. 679–684.
- [10] ———, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
- [11] ———, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NJ, 1961.
- [12] R. BELLMAN AND D. BLACKWELL, *On a Particular Non-Zero Sum Game*, RM-250, RAND Corp., Santa Monica, CA, 1949.
- [13] R. BELLMAN AND J. P. LA SALLE, *On Non-Zero Sum Games and Stochastic Processes*, RM-212, RAND Corp., Santa Monica, CA, 1949.
- [14] D. P. BERTSEKAS, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
- [15] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

- [16] F. J. BEUTLER AND K. W. ROSS, *Optimal policies for controlled Markov chain with a constraint*, J. Math. Anal. Appl., 112 (1985), pp. 236–256.
- [17] R. N. BHATTACHARYA AND M. MAJUMDAR, *Controlled semi-Markov model under long-run average rewards*, J. Statist. Plann. Inference, 22 (1989), pp. 223–242.
- [18] D. BLACKWELL, *Discrete dynamic programming*, Ann. Math. Statist., 33 (1962), pp. 719–726.
- [19] ———, *Discounted dynamic programming*, Ann. Math. Statist., 36 (1965), pp. 226–235.
- [20] V. S. BORKAR, *Controlled Markov chains and stochastic networks*, SIAM J. Control Optim., 21 (1983), pp. 652–666.
- [21] ———, *On minimum cost per unit time control of Markov chains*, SIAM J. Control Optim., 22 (1984), pp. 965–984.
- [22] ———, *Control of Markov chains with long-run average cost criterion*, in Stochastic Differential Systems, Stochastic Control Theory and Applications (W. Fleming and P. L. Lions, eds.), The IMA Volumes in Mathematics and Its Applications, Vol. 10, Springer-Verlag, Berlin, 1988, pp. 57–77.
- [23] ———, *A convex analytic approach to Markov decision processes*, Probab. Theory Related Fields, 78 (1988), pp. 583–602.
- [24] ———, *Control of Markov chains with long-run average cost criterion: The dynamic programming equations*, SIAM J. Control Optim., 27 (1989), pp. 642–657.
- [25] ———, *Controlled Markov chains with constraints*, Proceedings of the Workshop on Recent Advances in Modelling and Control of Stochastic Systems, Bangalore, India, January 1991, Indian Academy of Sciences, to appear.
- [26] ———, *Topics in Controlled Markov Chains*, Pitman Research Notes in Math. No. 240, Longman Scientific and Technical, Harlow, 1991.
- [27] ———, *Ergodic control of Markov chains with constraints — the general case*, preprint.
- [28] V. S. BORKAR AND M. K. GHOSH, *Ergodic and adaptive control of nearest neighbour motions*, Math. Control Signals Systems, 4 (1991), pp. 81–98.
- [29] L. D. BROWN AND R. PURVES, *Measurable selection of extrema*, Ann. Statist., 1 (1973), pp. 902–912.
- [30] R. CAVAZOS-CADENA, *Necessary and sufficient conditions for a bounded solution to the optimality equation in average reward Markov decision chains*, Systems Control Lett., 10 (1988), pp. 71–78.
- [31] ———, *Necessary conditions for the optimality equation in average reward Markov decision processes*, Appl. Math. Optim., 19 (1989), pp. 97–112.
- [32] ———, *Recent results on conditions for the existence of average optimal stationary policies*, Ann. Oper. Res., 28 (1991), pp. 3–26.
- [33] ———, *A counterexample on the optimality equation in Markov decision chains with the average cost criterion*, Systems Control Lett., 16 (1991), pp. 387–392.
- [34] R. CAVAZOS-CADENA AND L. I. SENNOTT, *Comparing recent assumptions for the existence of average optimal stationary policies*, Oper. Res. Lett., to appear.
- [35] R. YA. CHITASHVILI, *A controlled finite Markov chain with an arbitrary set of decisions*, Theory Probab. Appl., 20 (1975), pp. 839–846.
- [36] E. V. DENARDO, *A Markov decision problem*, in Mathematical Programming (T. C. Hu and S. M. Robinson, eds.), Academic Press, New York, 1973.
- [37] E. V. DENARDO AND B. L. FOX, *Multichain Markov renewal programs*, SIAM J. Appl. Math., 16 (1968), pp. 468–487.
- [38] C. DERMAN, *On sequential decisions and Markov chains*, Management Sci., 9 (1962), pp. 16–24.
- [39] ———, *Denumerable state Markov decision processes — average cost criterion*, Ann. Math. Statist., 37 (1966), pp. 1545–1553.
- [40] ———, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.
- [41] C. DERMAN AND M. KLEIN, *Some remarks on finite horizon Markovian decision models*, Oper. Res., 13 (1965), pp. 272–278.
- [42] C. DERMAN AND R. E. STRAUCH, *A note on memoryless rules for controlling sequential control processes*, Ann. Math. Statist., 37 (1966), pp. 276–278.
- [43] C. DERMAN AND A. F. VEINOTT, JR., *A solution to a countable system of equations arising in Markovian decision processes*, Ann. Math. Statist., 38 (1967), pp. 582–584.
- [44] ———, *Constrained Markov decision chains*, Management Sci., 19 (1972), pp. 389–390.
- [45] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
- [46] A. W. DRAKE, *Observation of a Markov Process Through a Noisy Channel*, Ph.D. thesis, Dept. of Electrical Engineering, MIT, Cambridge, MA, 1962.

- [47] L. DUBINS AND L. SAVAGE, *How to Gamble if You Must: Inequalities for Stochastic Processes*, McGraw-Hill, New York, 1965.
- [48] S. DURINOVIC, H. M. LEE, M. N. KATEHAKIS, AND J. A. FILAR, *Multiobjective Markov decision process with average reward criterion*, Large Scale Systems, 10 (1986), pp. 215–226.
- [49] A. DVORETZKY, J. KEIFER AND J. WOLFOWITZ, *The inventory problem*, Econometrica, 20 (1956), pp. 187–222; pp. 450–466.
- [50] E. B. DYNKIN, *Controlled random sequences*, Theory Probab. Appl., 10 (1965), pp. 1–14.
- [51] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.
- [52] A. FEDERGRUEN, A. HORDIJK, AND H. C. TIJMS, *Recurrent conditions in denumerable state Markov decision processes*, in Dynamic Programming and Its Applications, M. L. Puterman, ed., Academic Press, New York, 1978, pp. 3–22.
- [53] ———, *Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion*, Stochast. Process. Appl., 9 (1979), pp. 223–235.
- [54] A. FEDERGRUEN, P. J. SCHWEITZER, AND H. C. TIJMS, *Contraction mappings underlying undiscounted Markov decision problems*, J. Math. Anal. Appl., 65 (1978), pp. 711–730.
- [55] ———, *Denumerable undiscounted semi-Markov decision processes with unbounded rewards*, Math. Oper. Res., 8 (1983), pp. 298–313.
- [56] A. FEDERGRUEN AND H. C. TIJMS, *The optimality equation in average cost denumerable state semi-Markov decision problems, recurrence conditions and algorithms*, J. Appl. Probab., 15 (1978), pp. 356–373.
- [57] E. A. FEINBERG, *On controlled finite state Markov processes with compact control sets*, Theory Probab. Appl., 20 (1975), pp. 856–862.
- [58] ———, *The existence of a stationary ε -optimal policy for a finite Markov chain*, Theory Probab. Appl., 23 (1978), pp. 297–313.
- [59] ———, *An ε -optimal control of a finite Markov chain with an average reward criterion*, Theory Probab. Appl., 25 (1980), pp. 70–81.
- [60] ———, *Controlled Markov processes with arbitrary numerical criteria*, Theory Probab. Appl., 27 (1982), pp. 486–503.
- [61] E. FERNÁNDEZ-GAUCHERAND, *Controlled Markov Processes on the Infinite Planning Horizon: Optimal and Adaptive Control*, Ph.D. thesis, Electrical and Computer Engineering Dept., University of Texas at Austin, 1991.
- [62] E. FERNÁNDEZ-GAUCHERAND, A. ARAPOSTATHIS, AND S. I. MARCUS, *On partially observable Markov decision processes with an average cost criterion*, in Proc. 28th IEEE Conf. on Decision and Control, Tampa, FL, 1989, pp. 1267–1272.
- [63] ———, *On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes*, Ann. Oper. Res., 29 (1991), pp. 439–470.
- [64] ———, *Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes*, Systems Control Lett., 15 (1990), pp. 425–432.
- [65] E. FERNÁNDEZ-GAUCHERAND, M. K. GHOSH, AND S. I. MARCUS, *Controlled Markov processes on the infinite planning horizon with a weighted cost criterion*, Contribuciones en Probabilidad y Estadística Matemática, 3 (1992), pp. 145–162.
- [66] C. H. FINE, *A quality control model with learning effects*, Oper. Res., 36 (1988), pp. 437–444.
- [67] L. FISHER AND S. M. ROSS, *An example in denumerable decision processes*, Ann. Math. Statist., 39 (1968), pp. 674–675.
- [68] J. FLYNN, *Averaging versus discounting in dynamic programming: a counterexample*, Ann. Statist., 2 (1974), pp. 411–413.
- [69] ———, *Conditions for the equivalence of optimality criteria in dynamic programming*, Ann. Statist., 4 (1976), pp. 936–953.
- [70] ———, *On optimality criteria for dynamic programs with long finite horizons*, J. Math. Anal. Appl., 76 (1980), pp. 202–208.
- [71] N. FURUKAWA, *Markovian decision processes with compact action spaces*, Ann. Math. Statist., 43 (1972), pp. 1612–1622.
- [72] J.-P. GEORGIN, *Contrôle des chaînes de Markov sur des espaces arbitraires*, Ann. Inst. H. Poincaré Probab. Statist. Sect. B, 14 (1978), pp. 255–277.
- [73] ———, *Estimation et contrôle des chaînes de Markov sur des espaces arbitraires*, in Lecture Notes Math., 636, Springer-Verlag, Berlin, 1978, pp. 71–113.
- [74] M. K. GHOSH, *Ergodic and Adaptive Control of Markov Processes*, Ph.D. thesis, Indian Institute of Science, Bangalore, India, 1988.

- [75] M. K. GHOSH, *Markov decision processes with multiple costs*, Oper. Res. Lett., 9 (1990), pp. 257–260.
- [76] M. K. GHOSH AND S. I. MARCUS, *Ergodic control of Markov chains*, in Proc. 29th IEEE Conf. on Decision and Control, Honolulu, Hawaii, 1990, pp. 258–263.
- [77] ———, *On strong average optimality of Markov decision processes with unbounded costs*, Oper. Res. Lett., 11 (1992), pp. 99–104.
- [78] I. I. GIHMAN AND A. V. SKOROHOD, *Controlled Stochastic Processes*, Springer-Verlag, New York, 1979.
- [79] D. GILLETTE, *Stochastic games with zero stop probabilities*, Contributions to the Theory of Games, III, Annals of Math. Studies, 39, Princeton University Press, Princeton, NJ, 1957, pp. 71–187.
- [80] L. G. GUBENKO AND E. S. STATLAND, *On controlled, discrete-time Markov decision processes*, Theory Probab. Math. Statist., 7 (1975), pp. 47–61.
- [81] M. HAVIV AND M. L. PUTERMAN, *An improved algorithm for solving communicating average reward Markov decision processes*, Ann. Oper. Res., 29 (1991), pp. 229–242.
- [82] O. HERNÁNDEZ-LERMA, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.
- [83] ———, *Average optimality in dynamic programming on Borel spaces: Unbounded costs and controls*, preprint.
- [84] O. HERNÁNDEZ-LERMA, J. C. HENNET, AND J. B. LASSERRE, *Average cost Markov decision processes: optimality conditions*, J. Math. Anal. Appl., 158 (1991), pp. 396–406.
- [85] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Average cost optimal policies for Markov control processes with Borel state space and unbounded costs*, Systems Control Lett., 15 (1990), pp. 349–356.
- [86] O. HERNÁNDEZ-LERMA, R. MONTES-DE-OCA, AND R. CAVAZOS-CADENA, *Recurrence conditions for Markov decision processes with Borel state space: a survey*, Ann. Oper. Res., 29 (1991), pp. 29–46.
- [87] D. P. HEYMAN AND M. J. SOBEL, *Stochastic Models in Oper. Res., vol. II: Stochastic Optimization*, McGraw-Hill, New York, 1984.
- [88] C. J. HIMMELBERG, T. PARTHASARATHY, AND F. S. VAN VLECK, *Optimal plans for dynamic programming problems*, Math. Oper. Res., 1 (1976), pp. 390–394.
- [89] K. HINDERER, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameters*, Lecture Notes Oper. Res. Math. Systems, Vol. 33, Springer-Verlag, Berlin, 1970.
- [90] W. J. HOPP AND S. C. WU, *Multiaction maintenance under Markovian deterioration and incomplete information*, Naval Res. Logist. Quart., 35 (1988), pp. 447–462.
- [91] A. HORDIJK, *Dynamic Programming and Markov Potential Theory*, Math. Centre Tract, No. 51, Mathematisch Centrum, Amsterdam, 1974.
- [92] A. HORDIJK AND L. C. M. KALLENBERG, *Linear programming and Markov decision chains*, Management Sci., 25 (1979), pp. 352–362.
- [93] ———, *Constrained undiscounted stochastic dynamic programming*, Math. Oper. Res., 9 (1984), pp. 276–289.
- [94] A. HORDIJK AND M. L. PUTERMAN, *On the convergence of policy iteration in undiscounted finite state Markov processes; the unichain case*, Math. Oper. Res., 12 (1987), pp. 163–176.
- [95] R. HOWARD, *Dynamic Programming and Markov Decision Processes*, MIT Press, Cambridge, MA, 1960.
- [96] G. HÜBNER, *On the fixed points of the optimal reward operator in stochastic dynamic programming with discount factor greater than one*, Zeit. Angew. Math. Mech., 57 (1977), pp. 477–480.
- [97] L. C. M. KALLENBERG, *Linear Programming and Finite Markovian Control Problems*, Math. Centre Tract, No. 148, Mathematisch Centrum, Amsterdam, 1983.
- [98] S. KARLIN, *The structure of dynamic programming models*, Naval Res. Logist. Quart., 2 (1955), pp. 285–294.
- [99] D. KRASS, J. A. FILAR, AND S. SINHA, *A weighted Markov decision process*, Oper. Res., to appear.
- [100] N. V. KRYLOV, *Construction of an optimal strategy for a finite controlled chain*, Theory Probab., 10 (1965), pp. 45–54.
- [101] P. R. KUMAR, *Simultaneous identification and adaptive control of unknown systems over finite parameter sets*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 68–76.
- [102] ———, *A survey of some of results in stochastic adaptive control*, SIAM J. Control Optim., 23 (1985), pp. 329–380.

- [103] P. R. KUMAR AND P. VARAIYA, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
- [104] M. KURANO, *Markov decision processes with a Borel measurable cost function: the average case*, Math. Oper. Res., 11 (1986), pp. 309–320.
- [105] ———, *The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin*, SIAM J. Control Optim., 27 (1989), pp. 296–307.
- [106] ———, *Average Cost Markov Decision Processes under the Hypothesis of Doeblin*, Report No. 9, Dept. Mathematics, Faculty of Education, Chiba, Japan, 1989.
- [107] ———, *On Optimality Inequalities in Average Cost Markov Decision Processes with Doeblin's Conditions*, Report No. 1, Dept. Mathematics, Faculty of Education, Chiba, Japan, 1990.
- [108] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.
- [109] ———, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048.
- [110] B. L. LAMOND AND M. L. PUTERMAN, *Generalized inverses in discrete time Markov decision processes*, SIAM J. Math. Anal. Appl., 10 (1989), pp. 118–134.
- [111] A. LEIZAROWITZ, *Infinite horizon optimization for a finite state Markov chain*, SIAM J. Control Optim., 25 (1987), pp. 1601–1618.
- [112] G. DE LEVE, *Generalized Markov Decision Processes, Part I: Model and Method*, Math. Centre Tract, No. 3, Mathematisch Centrum, Amsterdam, 1964.
- [113] ———, *Generalized Markov Decision Processes, Part II: Probabilistic Background*, Math. Centre Tract, No. 4, Mathematisch Centrum, Amsterdam, 1964.
- [114] G. DE LEVE, A. FEDERGRUEN, AND H. C. TIJMS, *A general Markov decision method*, Adv. Appl. Probab., 9 (1977), pp. 296–335.
- [115] S. A. LIPPMAN, *Semi-Markov decision processes with unbounded rewards*, Management Sci., 19 (1973), pp. 717–731.
- [116] ———, *On dynamic programming with unbounded rewards*, Management Sci., 21 (1975), pp. 1225–1233.
- [117] M. LOËVE, *Probability Theory II*, Springer-Verlag, Berlin, 1978.
- [118] W. S. LOVEJOY, *Some monotonicity results for partially observed Markov decision processes*, Oper. Res., 35 (1987), pp. 736–743.
- [119] ———, *A survey of algorithmic methods for partially observed Markov decision processes*, Ann. Oper. Res., 28 (1991), pp. 47–66.
- [120] D.-J. MA, A. M. MAKOWSKI, AND A. SHWARTZ, *Estimation and optimal control for constrained Markov chains*, in Proc. 25th IEEE Conf. on Decision and Control, Athens, Greece, 1986, pp. 994–999.
- [121] A. MAITRA, *Dynamic Programming for Countable State Systems*, Ph.D. thesis, University of California, Berkeley, CA, 1964.
- [122] ———, *Dynamic programming for countable state systems*, Sankhyā Ser. A, 27 (1965), pp. 241–248.
- [123] ———, *Discounted dynamic programming on compact metric spaces*, Sankhyā Ser. A, 30 (1968), pp. 211–216.
- [124] P. MANDL, *Estimation and control in Markov chains*, Adv. Appl. Probab., 6 (1974), pp. 40–60.
- [125] A. MANNE, *Linear programming and sequential decisions*, Management Sci., 6 (1960), pp. 259–267.
- [126] A. MARTIN-LÖF, *Existence of a stationary control for a Markov chain maximizing the average reward*, Oper. Res., 15 (1967), pp. 866–871.
- [127] B. L. MILLER AND A. F. VEINOTT, JR., *Discrete dynamic programming with a small interest rate*, Ann. Math. Statist., 40 (1969), pp. 366–370.
- [128] G. E. MONAHAN, *A survey of partially observable Markov decision processes: theory, models, and algorithms*, Management Sci., 28 (1982), pp. 1–16.
- [129] P. NAIN AND K. W. ROSS, *Optimal priority assignment with hard constraint*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 883–888.
- [130] J. NEVEU, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, CA, 1965.
- [131] M. OHNISHI, H. KAWAI, AND H. MINE, *An optimal inspection and replacement policy under incomplete state information*, European J. Oper. Res., 27 (1986), pp. 117–128.
- [132] M. OHNISHI, H. MINE, AND H. KAWAI, *An optimal inspection and replacement policy under incomplete state information: average cost criterion*, in Stochastic Models in Reliability Theory (S. Osaki and Y. Hatoyama, eds.), Lect. Notes Econ. Math. Systems, Vol. 235, Springer-Verlag, Berlin, 1984, pp. 187–197.

- [133] S. OREY, *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand, London, 1971.
- [134] K. R. PARTHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.
- [135] L. K. PLATZMAN, *Finite Memory Estimation and Control of Finite Probabilistic Systems*, Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1977.
- [136] ———, *Optimal infinite horizon undiscounted control of finite probabilistic systems*, SIAM J. Control Optim., 18 (1980), pp. 362–380.
- [137] M. L. PUTERMAN, *Markov decision processes*, in *Handbooks in Operation Research and Management Science* (D. P. Heyman and M. J. Sobel, eds.), Vol. 2, North-Holland, Amsterdam, 1990, pp. 331–434.
- [138] D. RHENIUS, *Incomplete information in Markovian decision models*, Ann. Statist., 2 (1974), pp. 1327–1334.
- [139] U. RIEDER, *Measurable selection theorems for optimization problems*, Manuscripta Math., 24 (1978), pp. 115–131.
- [140] R. K. RITT AND L. I. SENNOTT, *Optimal stationary policies in general state Markov decision chains with finite action set*, Math. Oper. Res., to appear.
- [141] D. R. ROBINSON, *Markov decision chains with unbounded costs and applications to the control of queues*, Adv. Appl. Probab., 8 (1976), pp. 159–176.
- [142] ———, *Optimality conditions for a Markov decision chain with unbounded costs*, J. Appl. Probab., 17 (1980), pp. 996–1003.
- [143] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [144] Z. ROSBERG, P. VARAIYA, AND J. WALRAND, *Optimal control of service in tandem queues*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 600–610.
- [145] K. W. ROSS, *Randomized and past dependent policies for Markov decision processes with multiple constraints*, Oper. Res., 37 (1989), pp. 474–477.
- [146] K. W. ROSS AND R. VARADARAJAN, *Markov decision processes with sample path constraints: The communicating case*, Oper. Res., 37 (1989), pp. 780–790.
- [147] S. M. ROSS, *Non-discounted denumerable Markovian decision models*, Ann. Math. Statist., 39 (1968), pp. 412–423.
- [148] ———, *Arbitrary state Markovian decision processes*, Ann. Math. Statist., 39 (1968), pp. 2118–2122.
- [149] ———, *Quality control under Markovian deterioration*, Management Sci., 17 (1971), pp. 587–596.
- [150] ———, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.
- [151] Y. SAWARAGI AND T. YOSHIKAWA, *Discrete time Markovian decision processes with incomplete state observation*, Ann. Math. Statist., 41 (1970), pp. 78–86.
- [152] M. SCHÄL, *On continuous dynamic programming with discrete time parameters*, Z. Wahrsch. Verw. Geb., 21 (1972), pp. 279–288.
- [153] ———, *On dynamic programming: compactness of the space of policies*, Stochast. Process. Appl., 3 (1975), pp. 345–364.
- [154] ———, *Conditions for optimality in dynamic programming and for the limit of n -stage optimal policies to be optimal*, Z. Wahrsch. Verw. Geb., 32 (1975), pp. 179–196.
- [155] L. I. SENNOTT, *A new condition for the existence of optimal stationary policies in average cost Markov decision processes*, Oper. Res. Lett., 5 (1986), pp. 17–23.
- [156] ———, *Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs*, Oper. Res., 37 (1989), pp. 626–633.
- [157] ———, *Average cost semi-Markov decision processes and the control of queueing systems*, Probab. Engrg. Inform. Sci., 3 (1989), pp. 247–272.
- [158] ———, *The average cost optimality equation and critical number policies*, preprint.
- [159] R. F. SERFOZO, *An equivalence between continuous and discrete time Markov decision processes*, Oper. Res., 27 (1979), pp. 616–620.
- [160] L. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 1095–1100.
- [161] A. N. SHIRYAEV, *On the theory of decision functions and control by an observation process with incomplete data*, in *Selected Translations in Mathematical Statistics and Probability*, Vol. 6, American Mathematical Society, Providence, RI, 1966, pp. 162–188.
- [162] ———, *Some new results in the theory of controlled random sequences*, in *Selected Translations in Mathematical Statistics and Probability*, Vol. 8, American Mathematical Society, Providence, RI, 1970, pp. 49–130.

- [163] A. N. SHIRYAEV, *On Markov sufficient statistics in non-additive Bayes problems of sequential analysis*, Theory Probab. Appl., 9 (1964), pp. 604–618.
- [164] S. E. SHREVE AND D. P. BERTSEKAS, *Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control*, SIAM J. Control Optim., 16 (1978), pp. 953–978.
- [165] ———, *Dynamic programming in Borel spaces*, in Dynamic Programming and Its Applications, M. L. Puterman, ed., Academic Press, New York, 1978, pp. 115–130.
- [166] A. SHWARTZ AND A. M. MAKOWSKI, *An optimal adaptive scheme for two competing queues with constraints*, in Analysis and Optimization of Systems (A. Bensoussan and J. L. Lions, eds.), Lecture Notes on Control and Information Sciences, Springer-Verlag, Berlin, 1986.
- [167] ———, *On the Poisson Equation for Markov Chains*, Report No. EE-646, Faculty of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel, 1987.
- [168] ———, *Comparing policies in Markov decision processes: Mandl's lemma revisited*, Math. Oper. Res., 15 (1990), pp. 155–174.
- [169] R. D. SMALLWOOD AND E. J. SONDIK, *The optimal control of partially observable Markov process over a finite horizon*, Oper. Res., 21 (1973), pp. 1071–1088.
- [170] E. J. SONDIK, *The Optimal Control of Partially Observable Markov Processes*, Ph.D. thesis, Electrical Engineering Dept., Stanford University, Stanford, CA, 1971.
- [171] ———, *The optimal control of partially observable Markov decision problems over the infinite horizon: Discounted costs*, Oper. Res., 26 (1978), pp. 282–304.
- [172] S. S. STIDHAM JR. AND R. R. WEBER, *Monotonic and insensitive optimal policies for control of queues with unbounded costs*, Oper. Res., 87 (1989), pp. 611–625.
- [173] R. E. STRAUCH, *Negative dynamic programming*, Ann. Math. Statist., 37 (1966), pp. 871–890.
- [174] C. STRIEBEL, *Sufficient statistics in the optimum control of stochastic systems*, J. Math. Anal. Appl., 12 (1965), pp. 576–593.
- [175] ———, *Optimal Control of Discrete Time Stochastic Systems*, Lecture Notes Econom. Math. Systems, Vol. 110, Springer-Verlag, Berlin, 1975.
- [176] R. SZNADJER AND J. A. FILAR, *Some comments on a theorem of Hardy and Littlewood*, J. Optim. Theory Appl., 75 (1992), to appear.
- [177] H. M. TAYLOR, *Markovian sequential replacement processes*, Ann. Math. Statist., 38 (1965), pp. 1677–1694.
- [178] L. C. THOMAS, *Connectedness conditions for denumerable state Markov decision processes*, in Recent Developments in Markov Decision Processes (R. Hartley, L. C. Thomas, and D. F. White, eds.), Academic Press, New York, 1980, pp. 181–204.
- [179] H. C. TIJMS, *On Dynamic Programming with Arbitrary State Space, Compact Action Space and the Average Reward as Criterion*, Report BW 55/75, Mathematisch Centrum, Amsterdam, 1975.
- [180] ———, *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley, Chichester, UK, 1986.
- [181] J. VAN DER WAL AND J. WESSELS, *Markov decision processes*, Statist. Neerlandica, 39 (1985), pp. 219–233.
- [182] A. F. VEINOTT, *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40 (1969), pp. 1635–1660.
- [183] O. V. VISKOV AND A. N. SHIRYAEV, *On controls leading to optimal stationary models*, Trudy Mat. Inst. Steklov, 71 (1964), pp. 35–45. (In Russian.)
- [184] D. H. WAGNER, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859–903.
- [185] H. M. WAGNER, *On the optimality of pure strategies*, Management Sci., 6 (1960), pp. 268–269.
- [186] A. WALD, *Sequential Analysis*, John Wiley, New York, 1947.
- [187] ———, *Statistical Decision Functions*, John Wiley, New York, 1950.
- [188] R. WANG, *Optimal replacement policy with unobservables states*, J. Appl. Probab., 14 (1977), pp. 340–348.
- [189] ———, *Computing optimal quality control policies — two actions*, J. Appl. Probab., 14 (1977), pp. 826–832.
- [190] R. R. WEBER AND S. S. STIDHAM JR., *Optimal control of service rates in networks of queues*, Adv. Appl. Probab., 15 (1987), pp. 202–218.
- [191] C. C. WHITE, *A Markov quality control process subject to partial observation*, Management Sci., 23 (1977), pp. 843–852.
- [192] ———, *Optimal inspection and repair of a production process subject to deterioration*, J. Oper. Res. Soc., 29 (1978), pp. 235–243.

- [193] C. C. WHITE, *Bounds on optimal cost for a replacement problem with partial observation*, Naval Res. Logist. Quart., 26 (1979), pp. 415–422.
- [194] ———, *Optimal control — limit strategies for a partially observed replacement problem*, Internat. J. Systems Sci., 10 (1979), pp. 321–331.
- [195] ———, *Monotone control laws for noisy, countable-state Markov chains*, European J. Oper. Res., 5 (1980), pp. 124–132.
- [196] C. C. WHITE AND D. J. WHITE, *Markov decision processes*, European J. Oper. Res., 39 (1989), pp. 1–16.
- [197] D. J. WHITE, *Dynamic programming of Markov chains and the method of successive approximations*, J. Math. Anal. Appl., 6 (1963), pp. 373–376.
- [198] ———, *Mean, variance, and probabilistic criteria in finite Markov decision processes: A review*, J. Optim. Theory Appl., 56 (1988), pp. 1–29.
- [199] P. WHITTLE, *Sequential decision processes with essential unobservables*, Adv. Appl. Probab., 1 (1969), pp. 271–287.
- [200] ———, *Optimization over Time: Dynamic Programming and stochastic control*, II, John Wiley, Chichester, UK, 1983.
- [201] J. WIJNGAARD, *Stationary Markovian decision problems and perturbation theory of quasicompact linear operators*, Math. Oper. Res., 2 (1977), pp. 91–102.
- [202] ———, *Existence of average optimal strategies in Markovian decision problems with strictly unbounded costs*, in Dynamic Programming and Its Applications, M. L. Puterman, ed., Academic Press, New York, 1978, pp. 369–386.
- [203] K. YAMADA, *Duality theorem in Markovian decision problems*, J. Math. Anal. Appl., 50 (1975), pp. 579–595.
- [204] A. A. YUSHKEVICH, *On a class of strategies in general Markov decision models*, Theory Probab. Appl., 18 (1973), pp. 777–779.
- [205] ———, *Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces*, Theory Probab. Appl., 21 (1976), pp. 153–158.
- [206] ———, *On reducing a jump controllable Markov model to a model with discrete time*, Theory Probab. Appl., 25 (1980), pp. 58–59.
- [207] A. A. YUSHKEVICH AND R. YA. CHITASHVILI, *Controlled random sequences and Markov chains*, Russian Math. Surveys, 37 (1982), pp. 239–274.
- [208] H. ZIJM, *The optimality equations in multichain denumerable Markov decision processes with average cost criterion: The bounded cost case*, Statist. Decisions, 3 (1985), pp. 143–165.

FLEMING–VIOT PROCESSES IN POPULATION GENETICS*

S. N. ETHIER[†] AND THOMAS G. KURTZ[‡]

This paper is dedicated to Wendell Fleming on the occasion of his 65th birthday.

Abstract. Fleming and Viot [*Indiana Univ. Math. J.*, 28 (1979), pp. 817–843] introduced a class of probability-measure-valued diffusion processes that has attracted the interest of both pure and applied probabilists. This paper surveys the subject of Fleming–Viot processes as it relates to population genetics. Topics include:

1. Introduction.
2. Some measure-valued Markov chains.
- 2.1. A diploid model.
- 2.2. The Wright–Fisher model.
- 2.3. A Moran model.
3. The Fleming–Viot process: characterization.
4. Convergence.
5. Ergodicity.
6. An infinite particle system.
7. Bounded mutation operators.
8. Reversibility.
9. Examples.
- 9.1. Continuous-state stepwise-mutation model.
- 9.2. Infinitely-many-neutral-alleles model.
- 9.3. Infinitely-many-neutral-alleles model with ages.
- 9.4. Two-locus model with recombination.
- 9.5. n -locus model with gene conversion.
- 9.6. Infinitely-many-sites model without recombination.
- 9.7. Infinitely-many-neutral-alleles model with allelic genealogies.

Key words. measure-valued Markov chain, measure-valued diffusion process, martingale problem, convergence in distribution, ergodicity, infinite particle system, reversibility, infinitely-many-neutral-alleles model

AMS(MOS) subject classifications. 60G57, 60J60, 92D15

1. Introduction. The neutral diffusion model in population genetics, in which each individual is of some “type” and the set E of types is finite, has state space

$$(1.1) \quad \Delta_E = \left\{ (p_i)_{i \in E} \in [0, 1]^E : \sum_{i \in E} p_i = 1 \right\},$$

where p_i denotes the proportion of the population that is of type i . Its generator is

$$(1.2) \quad L = \frac{1}{2} \sum_{i, j \in E} p_i (\delta_{ij} - p_j) \frac{\partial^2}{\partial p_i \partial p_j} + \sum_{j \in E} \left(\sum_{i \in E} q_{ij} p_i \right) \frac{\partial}{\partial p_j},$$

where $(q_{ij})_{i, j \in E}$ is the infinitesimal matrix for a Markov process in E ; for $i \neq j$, q_{ij} is interpreted as the intensity of a mutation from type i to type j .

*Received by the editors January 3, 1992; accepted for publication (in revised form) July 7, 1992. This research was supported in part by National Science Foundation grants DMS-8901464 and DMS-9102925.

[†]Department of Mathematics, University of Utah, Salt Lake City, Utah 84112.

[‡]Department of Mathematics, University of Wisconsin–Madison, Madison, Wisconsin 53706.

Except for some technical requirements on $(q_{ij})_{i,j \in E}$, the same description is valid when E is countably infinite. One such example is Ohta and Kimura's (1973) stepwise-mutation model, in which $E = \mathbf{Z}$ and

$$(1.3) \quad q_{ij} = \begin{cases} \frac{1}{2}\theta & \text{if } j = i \pm 1 \\ -\theta & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

for some $\theta > 0$.

The case in which E is uncountably infinite, however, requires a different approach. The key idea, due to Fleming and Viot (1979), is to topologize E and replace Δ_E by $\mathcal{P}(E)$, the set of Borel probability measures on E with the topology of weak convergence. Then (1.2) becomes

$$(1.4) \quad (\mathcal{L}\varphi)(\mu) = \frac{1}{2} \int_E \int_E \mu(dx)(\delta_x(dy) - \mu(dy)) \frac{\delta^2 \varphi(\mu)}{\delta \mu(x) \delta \mu(y)} \\ + \int_E \mu(dx) A \left(\frac{\delta \varphi(\mu)}{\delta \mu(\cdot)} \right) (x),$$

where $\delta \varphi(\mu)/\delta \mu(x) = \lim_{\varepsilon \rightarrow 0+} \varepsilon^{-1} \{\varphi(\mu + \varepsilon \delta_x) - \varphi(\mu)\}$ and A is the generator for a Markov process in E . Here $\delta_x \in \mathcal{P}(E)$ denotes the unit mass at $x \in E$. The resulting probability-measure-valued diffusion process is referred to as a *Fleming-Viot process*. E is called the *type space* and A is known as the *mutation operator*. Terms corresponding to recombination and selection can also be included in (1.4); see (3.12) below.

It may reasonably be asked why an uncountable type space is needed. There are at least three reasons, and they are illustrated by the following three examples from population genetics, about which we will have more to say in §9.

(a) Continuous-state stepwise-mutation model (Ohta and Kimura (1973), Fleming and Viot (1979)). Here $E = \mathbf{R}$ and

$$(1.5) \quad (Af)(x) = \frac{1}{2}\theta f''(x),$$

which arises as a limit of the case $E = \mathbf{Z}$ and (1.3) after a suitable rescaling. The idea is that the type of an individual could be a quantitative characteristic measured on a continuum, thereby requiring an uncountable type space.

(b) Infinitely-many-neutral-alleles model (Kimura and Crow (1964), Watterson (1976), Ethier and Kurtz (1981), (1986), (1987)). Here E is arbitrary (except as noted below) and

$$(1.6) \quad (Af)(x) = \frac{1}{2}\theta \int_E (f(\xi) - f(x)) P(x, d\xi).$$

Mutations occur with intensity $\frac{1}{2}\theta$, and the type of a mutant offspring of a type x parent is distributed according to the one-step transition function $P(x, d\xi)$. The basic assumption of the model is that every mutant is of a new type, which requires that $P(x, \cdot)$ be nonatomic (i.e., have no atoms) for each $x \in E$. For this we need E uncountable.

(c) Infinitely-many-sites model without recombination (Kimura (1969), (1971), Watterson (1975), Ethier and Griffiths (1987)). Here $E = [0, 1]^{\mathbf{Z}_+}$ and

$$(1.7) \quad (Af)(\mathbf{x}) = \frac{1}{2}\theta \int_0^1 (f(\xi, \mathbf{x}) - f(\mathbf{x})) d\xi.$$

The interpretation is that $[0,1]$ is the set of sites on the chromosome, and an individual is of type $\mathbf{x} = (x_0, x_1, \dots) \in E$ if x_0, x_1, \dots is the sequence of sites at which mutations have occurred in the line of descent of that individual. Note that, even if $[0,1]$ were replaced by a finite set (with at least two elements), E would still be uncountable. The idea of using the type space to keep track of certain aspects of the history of the process is a very useful one.

As the work of Dawson, Dynkin, Le Gall, Perkins, Shiga, and others suggests, the subject of measure-valued diffusion processes has become an active and important branch of probability theory in recent years. The class of Fleming-Viot processes is surely one of the two most studied classes of measure-valued diffusions. It is our aim here to survey the subject of Fleming-Viot processes. Emphasis will be on results that relate to diffusion models in population genetics. This was, after all, the original motivation of Fleming and Viot (1979).

We assume throughout that (E, r) is a complete separable metric space. However, in most applications, E is compact or (as in example (a) above) locally compact. Of course, the locally compact case can be reduced to the compact case by a one-point compactification. Thus, we assume compactness of E whenever it is convenient to do so.

2. Some measure-valued Markov chains. The Fleming-Viot process arises most naturally as the limit in distribution of certain sequences of Markov chains occurring in population genetics. Here we describe three such models.

2.1. A diploid model. We define an equivalence relation \sim on $E^2 = E \times E$ as follows: $(x, y) \sim (z, w)$ if $(x, y) = (z, w)$ or $(x, y) = (w, z)$. The quotient set $E^{(2)}$ of all equivalence classes can be regarded as the set of all unordered pairs $\{x, y\}$ of elements of E . In a diploid population, chromosomes occur in pairs, and so an individual is described for our purposes by its genotype, which is an element of $E^{(2)}$. The formula

$$(2.1.1) \quad r^{(2)}(\{x, y\}, \{z, w\}) = (r(x, z) + r(y, w)) \wedge (r(x, w) + r(y, z))$$

defines a metric for $E^{(2)}$, and the quotient map $\rho : E^2 \mapsto E^{(2)}$ given by $\rho(x, y) = \{x, y\}$ is continuous. It can be shown that, if f is symmetric and Borel measurable on E^2 , then the function g on $E^{(2)}$ defined by $g(\{x, y\}) = f(x, y) = f(y, x)$ is Borel measurable.

Given $\mu \in \mathcal{P}(E^{(2)})$, we define its *symmetrization* $\hat{\mu} \in \mathcal{P}(E^2)$ by

$$(2.1.2) \quad \hat{\mu}(\Gamma) = \int_{E^{(2)}} \frac{1}{2} (\delta_{(x,y)}(\Gamma) + \delta_{(y,x)}(\Gamma)) \mu(d\{x, y\}).$$

For example, if $\mu = \delta_{\{x,y\}}$, then $\hat{\mu} = \frac{1}{2}(\delta_{(x,y)} + \delta_{(y,x)})$. Note that $\hat{\mu}\rho^{-1} = \mu$, so μ can be recovered from its symmetrization. We say that $\mu \in \mathcal{P}(E^{(2)})$ is in *Hardy-Weinberg form* if $\hat{\mu}$ is a product measure, that is, if $\hat{\mu} = \hat{\mu}\pi^{-1} \times \hat{\mu}\pi^{-1} = (\hat{\mu}\pi^{-1})^2$, where π is the projection of E^2 onto its first coordinate.

This model, as well as the other two, depends on certain “parameters,” which we introduce here for all three models. For each positive integer M , let w_M be a positive, symmetric, bounded, Borel function on E^2 , let $R_M((x, y), dx' \times dy')$ be a one-step transition function on $E^2 \times \mathcal{B}(E^2)$ satisfying

$$(2.1.3) \quad R_M((x, y), dx' \times dy') = R_M((y, x), dy' \times dx'),$$

and let $Q_M(x, dx')$ be a one-step transition function on $E \times \mathcal{B}(E)$. The functions w_M , R_M , and Q_M involve selection, recombination, and mutation, respectively.

Let $N \geq 1$ be the diploid population size. It will be convenient to define the mapping $\eta_N : (E^{(2)})^N \mapsto \mathcal{P}(E^{(2)})$ by letting $\eta_N(\{x_1, y_1\}, \dots, \{x_N, y_N\})$ be the empirical distribution determined by the (not necessarily distinct) points $\{x_1, y_1\}, \dots, \{x_N, y_N\} \in E^{(2)}$:

$$(2.1.4) \quad \eta_N(\{x_1, y_1\}, \dots, \{x_N, y_N\}) = N^{-1}(\delta_{\{x_1, y_1\}} + \dots + \delta_{\{x_N, y_N\}}).$$

The state space for the model is

$$(2.1.5) \quad \mathcal{P}_N(E^{(2)}) \equiv \eta_N((E^{(2)})^N) \subset \mathcal{P}(E^{(2)}).$$

Given $\mu \in \mathcal{P}(E^{(2)})$, we define $\mu_{2N}^*, \mu_{2N}^{**}, \mu_{2N}^{***} \in \mathcal{P}(E^{(2)})$ by

$$(2.1.6) \quad \hat{\mu}_{2N}^*(dx \times dy) = w_{2N}(x, y)(\hat{\mu}\pi^{-1})^2(dx \times dy) / \langle w_{2N}, (\hat{\mu}\pi^{-1})^2 \rangle,$$

$$(2.1.7) \quad \hat{\mu}_{2N}^{**}(dx' \times dy') = \int_{E^2} R_{2N}((x, y), dx' \times dy') \hat{\mu}_{2N}^*(dx \times dy),$$

$$(2.1.8) \quad \hat{\mu}_{2N}^{***}(dx' \times dy') = \int_{E^2} Q_{2N}(x, dx') Q_{2N}(y, dy') \hat{\mu}_{2N}^{**}(dx \times dy);$$

in general, $\langle f, \mu \rangle = \int_E f d\mu$ for $f \in B(E)$ and $\mu \in \mathcal{P}(E)$. (The condition (2.1.3) ensures that $\hat{\mu}_{2N}^*$ is the symmetrization of a measure in $\mathcal{P}(E^{(2)})$.) The Markov chain has one-step transition function $P_N(\mu, d\nu)$ on $\mathcal{P}_N(E^{(2)}) \times \mathcal{B}(\mathcal{P}_N(E^{(2)}))$ given by

$$(2.1.9) \quad P_N(\mu, \cdot) = \int_{(E^{(2)})^N} (\mu_{2N}^{***})^N(d\{x_1, y_1\} \times \dots \times d\{x_N, y_N\}) \delta_{\eta_N(\{x_1, y_1\}, \dots, \{x_N, y_N\})}(\cdot).$$

(A measure raised to the N th power denotes its N -fold product measure.) The model is a slight generalization of a model of Ethier and Kurtz (1987), which in turn generalizes a model of Ethier and Nagylaki (1980).

The interpretation is as follows. If $\mu \in \mathcal{P}_N(E^{(2)})$ is the empirical distribution of the N genotypes in the parent generation, then the empirical distribution of the N genotypes in the offspring generation is determined from μ in the four steps (2.1.6)–(2.1.9), corresponding, respectively, to reproduction and selection, recombination, mutation, and regulation (random sampling). In particular, (2.1.6) implicitly assumes that an infinite number of zygotes are produced in Hardy–Weinberg form as the initial step in the life cycle. See Nagylaki (1990) for a detailed discussion of the original formulation of the model.

2.2. The Wright–Fisher model. This model is mathematically simpler than the preceding one, but less reasonable biologically. In a diploid population of size N , there are $M = 2N$ gametes. Here we consider only the empirical distribution of the gametic types. It is unnecessary to require that M be even, so let M be a positive integer, and define $\eta_M : E^M \mapsto \mathcal{P}(E)$ as in (2.1.4):

$$(2.2.1) \quad \eta_M(x_1, \dots, x_M) = M^{-1}(\delta_{x_1} + \dots + \delta_{x_M}).$$

The state space for this model is

$$(2.2.2) \quad \mathcal{P}_M(E) \equiv \eta_M(E^M) \subset \mathcal{P}(E).$$

Given $\mu \in \mathcal{P}(E)$, we define $\mu_M^* \in \mathcal{P}(E^2)$ and $\mu_M^{**}, \mu_M^{***} \in \mathcal{P}(E)$ by

$$(2.2.3) \quad \mu_M^*(dx \times dy) = w_M(x, y) \mu^2(dx \times dy) / \langle w_M, \mu^2 \rangle,$$

$$(2.2.4) \quad \mu_M^{**}(dx') = \int_{E^2} R_M((x, y), dx' \times E) \mu_M^*(dx \times dy),$$

$$(2.2.5) \quad \mu_M^{***}(dx') = \int_E Q_M(x, dx') \mu_M^{**}(dx),$$

where w_M , R_M , and Q_M are as in §2.1. The Markov chain has one-step transition function $P_M(\mu, d\nu)$ on $\mathcal{P}_M(E) \times \mathcal{B}(\mathcal{P}_M(E))$ given by

$$(2.2.6) \quad P_M(\mu, \cdot) = \int_{E^M} (\mu_M^{***})^M(dx_1 \times \cdots \times dx_M) \delta_{\eta_M(x_1, \dots, x_M)}(\cdot).$$

For future reference, we note that $P_M(\mu, d\nu)$ can be extended to $\mathcal{P}(E) \times \mathcal{B}(\mathcal{P}_M(E))$.

The present formulation of the model is from Ethier and Kurtz (1993a), and the interpretation is similar to that in §2.1. The Wright–Fisher model has some characteristics of a diploid model (e.g., (2.2.3) and (2.2.4)) and some characteristics of a haploid model (e.g., (2.2.6)). It is, nevertheless, the most widely used discrete stochastic model in population genetics.

We note that, if we define $\gamma : \mathcal{P}(E^{(2)}) \mapsto \mathcal{P}(E)$ by $\gamma(\mu) = \hat{\mu}\pi^{-1}$, the image of the diploid model of §2.1 under γ is not generally the same as the Wright–Fisher model (with $M = 2N$). It is the same if the fitness function is multiplicative (i.e., $w_{2N}(x, y) = v_{2N}(x)v_{2N}(y)$) and recombination is absent (i.e., $R_M((x, y), \cdot) = \delta_{(x, y)}(\cdot)$), for then $\mu_{2N}^{***} \in \mathcal{P}(E^{(2)})$ is in Hardy–Weinberg form for all $\mu \in \mathcal{P}_{2N}(E^{(2)})$.

2.3. A Moran model. There are many variants of the model originated by Moran (1958). We describe a particularly simple one that allows us to use the notation introduced earlier. Unlike in the two preceding models, generations are overlapping. Transitions involve the death of an individual and the birth of another. Let $M \geq 1$ be the number of gametes in the population. The Markov chain has state space $\mathcal{P}_M(E)$ and one-step transition function $P_M(\mu, d\nu)$ on $\mathcal{P}_M(E) \times \mathcal{B}(\mathcal{P}_M(E))$ given by

$$(2.3.1) \quad P_M(\mu, \cdot) = \int_E \mu(dx) \int_E \mu_M^{***}(dx') \delta_{\mu - M^{-1}\delta_x + M^{-1}\delta_{x'}}(\cdot),$$

where $\mu_M^{***} \in \mathcal{P}(E)$ is defined in terms of $\mu \in \mathcal{P}(E)$ by (2.2.3)–(2.2.5). The interpretation is clear.

3. The Fleming–Viot process: characterization. Let $\{\nu_\tau^{(M)}, \tau \in \mathbf{Z}_+\}$ ($M = 1, 2, \dots$) be a sequence of Wright–Fisher models as described in §2.2. Under suitable conditions on E and the sequences $\{w_M\}$, $\{R_M\}$, and $\{Q_M\}$, and assuming weak convergence of initial distributions, it can be shown (see §4) that

$$(3.1) \quad \{\nu_{[Mt]}^{(M)}, t \geq 0\} \Rightarrow \{\mu_t, t \geq 0\} \quad \text{in } D_{\mathcal{P}(E)}[0, \infty) \quad \text{as } M \rightarrow \infty,$$

where $\{\mu_t, t \geq 0\}$ is a diffusion process in $\mathcal{P}(E)$. In this section we identify and characterize the limiting diffusion.

Let \mathcal{L}_M denote the discrete generator for the M th rescaled Markov chain:

$$(3.2) \quad (\mathcal{L}_M \varphi)(\mu) = M \int_{\mathcal{P}_M(E)} (\varphi(\nu) - \varphi(\mu)) P_M(\mu, d\nu),$$

where P_M is given by (2.2.6); we regard (3.2) as being defined on all of $\mathcal{P}(E)$, not just on $\mathcal{P}_M(E)$. We want to find sufficient conditions for the limit of (3.2) to exist as $M \rightarrow \infty$. We initially restrict our attention to test functions φ of the form

$$(3.3) \quad \varphi(\mu) = \langle f_1, \mu \rangle \cdots \langle f_k, \mu \rangle,$$

where $k \geq 1$ and $f_1, \dots, f_k \in B(E)$; recall that $\langle f, \mu \rangle = \int_E f d\mu$. By (2.2.6),

$$(3.4) \quad \begin{aligned} & (\mathcal{L}_M \varphi)(\mu) \\ &= M \int_{\mathcal{P}_M(E)} \left\{ \prod_{i=1}^k \langle f_i, \nu \rangle - \prod_{i=1}^k \langle f_i, \mu \rangle \right\} P_M(\mu, d\nu) \\ &= M \left\{ \int_{E^M} \prod_{i=1}^k \langle f_i, \eta_M(x_1, \dots, x_M) \rangle (\mu_M^{***})^M(dx_1 \times \cdots \times dx_M) - \prod_{i=1}^k \langle f_i, \mu \rangle \right\} \\ &= M \left\{ M^{-k} \int_{E^M} \sum_{j_1=1}^M \cdots \sum_{j_k=1}^M f_1(x_{j_1}) \cdots f_k(x_{j_k}) (\mu_M^{***})^M(dx_1 \times \cdots \times dx_M) \right. \\ & \quad \left. - \prod_{i=1}^k \langle f_i, \mu \rangle \right\} \\ &= M \left\{ O(M^{-2}) + M^{-k} \frac{M!}{(M-k+1)!} \sum_{1 \leq i < j \leq k} \langle f_i f_j, \mu_M^{***} \rangle \prod_{l: l \neq i, j} \langle f_l, \mu_M^{***} \rangle \right. \\ & \quad \left. + M^{-k} \frac{M!}{(M-k)!} \prod_{i=1}^k \langle f_i, \mu_M^{***} \rangle - \prod_{i=1}^k \langle f_i, \mu \rangle \right\} \\ &= \sum_{1 \leq i < j \leq k} (\langle f_i f_j, \mu_M^{***} \rangle - \langle f_i, \mu_M^{***} \rangle \langle f_j, \mu_M^{***} \rangle) \prod_{l: l \neq i, j} \langle f_l, \mu_M^{***} \rangle \\ & \quad + \sum_{i=1}^k M (\langle f_i, \mu_M^{***} \rangle - \langle f_i, \mu \rangle) \prod_{l: l < i} \langle f_l, \mu \rangle \prod_{l: l > i} \langle f_l, \mu_M^{***} \rangle + O(M^{-1}), \end{aligned}$$

uniformly in $\mu \in \mathcal{P}(E)$. This calculation is from Kurtz (1981).

To ensure that this converges, we assume the existence of $\sigma \in B_{\text{sym}}(E^2)$ (the *selection intensity function*), a bounded linear transformation B from $B(E)$ to $B(E^2)$ (the *recombination operator*) of the form

$$(3.5) \quad (Bf)(x, y) = \alpha \int_E (f(x') - f(x)) R((x, y), dx'),$$

where $\alpha \geq 0$ (the *recombination intensity*) and $R((x, y), dx')$ is a one-step transition function on $E^2 \times \mathcal{B}(E)$, and a possibly unbounded linear operator A on $B(E)$ (the *mutation operator*, defined only on a subspace $\mathcal{D}(A)$) such that

$$(3.6) \quad w_M(x, y) = 1 + M^{-1} \sigma(x, y) + o(M^{-1}),$$

$$(3.7) \quad \int_E f(x') R_M((x, y), dx' \times E) = f(x) + M^{-1} (Bf)(x, y) + o(M^{-1}),$$

$$(3.8) \quad \int_E f(x') Q_M(x, dx') = f(x) + M^{-1} (Af)(x) + o(M^{-1}),$$

for all $f \in B(E)$ and $f \in \mathcal{D}(A)$, respectively, uniformly in $x, y \in E$. This then implies that

$$\begin{aligned}
 (3.9) \quad & \langle f, \mu_M^{***} \rangle \\
 &= \langle f + M^{-1}Af, \mu_M^{**} \rangle + o(M^{-1}) \\
 &= \langle (f + M^{-1}Af) \circ \pi + M^{-1}B(f + M^{-1}Af), \mu_M^* \rangle + o(M^{-1}) \\
 &= \frac{\langle ((f + M^{-1}Af) \circ \pi + M^{-1}Bf)(1 + M^{-1}\sigma), \mu^2 \rangle}{\langle 1 + M^{-1}\sigma, \mu^2 \rangle} + o(M^{-1}) \\
 &= \langle f, \mu \rangle + M^{-1}\{\langle Af, \mu \rangle + \langle Bf, \mu^2 \rangle + \langle (f \circ \pi)\sigma, \mu^2 \rangle - \langle f, \mu \rangle \langle \sigma, \mu^2 \rangle\} + o(M^{-1})
 \end{aligned}$$

for all $f \in \mathcal{D}(A)$, uniformly in $\mu \in \mathcal{P}(E)$, where π is the projection of E^2 onto its first coordinate.

Thus, if for $\varphi \in B(\mathcal{P}(E))$ of the form (3.3) with $k \geq 1$ and $f_1, \dots, f_k \in \mathcal{D}(A)$ we define

$$\begin{aligned}
 (3.10) \quad (\mathcal{L}\varphi)(\mu) &= \sum_{1 \leq i < j \leq k} (\langle f_i f_j, \mu \rangle - \langle f_i, \mu \rangle \langle f_j, \mu \rangle) \prod_{l: l \neq i, j} \langle f_l, \mu \rangle \\
 &\quad + \sum_{i=1}^k \{\langle Af_i, \mu \rangle + \langle Bf_i, \mu^2 \rangle\} \prod_{l: l \neq i} \langle f_l, \mu \rangle \\
 &\quad + \sum_{i=1}^k \{\langle (f_i \circ \pi)\sigma, \mu^2 \rangle - \langle f_i, \mu \rangle \langle \sigma, \mu^2 \rangle\} \prod_{l: l \neq i} \langle f_l, \mu \rangle,
 \end{aligned}$$

then, assuming that $\overline{\mathcal{D}(A)}$ is an algebra, (3.4) and (3.9) imply that

$$(3.11) \quad (\mathcal{L}_M \varphi)(\mu) = (\mathcal{L}\varphi)(\mu) + o(1),$$

uniformly in $\mu \in \mathcal{P}(E)$. More generally, we define \mathcal{L} by

$$\begin{aligned}
 (3.12) \quad (\mathcal{L}\varphi)(\mu) &= \frac{1}{2} \int_E \int_E \mu(dx)(\delta_x(dy) - \mu(dy)) \frac{\delta^2 \varphi(\mu)}{\delta \mu(x) \delta \mu(y)} \\
 &\quad + \int_E \mu(dx) A\left(\frac{\delta \varphi(\mu)}{\delta \mu(\cdot)}\right)(x) + \int_E \int_E \mu(dx) \mu(dy) B\left(\frac{\delta \varphi(\mu)}{\delta \mu(\cdot)}\right)(x, y) \\
 &\quad + \int_E \int_E \mu(dx) \mu(dy) (\sigma(x, y) - \langle \sigma, \mu^2 \rangle) \frac{\delta \varphi(\mu)}{\delta \mu(x)},
 \end{aligned}$$

where $\delta \varphi(\mu) / \delta \mu(x) = \lim_{\varepsilon \rightarrow 0+} \varepsilon^{-1} \{\varphi(\mu + \varepsilon \delta_x) - \varphi(\mu)\}$, and we take $\mathcal{D}(\mathcal{L})$ to be the set of all $\varphi \in B(\mathcal{P}(E))$ of the form

$$(3.13) \quad \varphi(\mu) = F(\langle f_1, \mu \rangle, \dots, \langle f_k, \mu \rangle) = F(\langle \mathbf{f}, \mu \rangle),$$

where $k \geq 1$, $f_1, \dots, f_k \in \mathcal{D}(A)$, and $F \in C^2(\mathbf{R}^k)$. For such φ ,

$$\begin{aligned}
 (3.14) \quad (\mathcal{L}\varphi)(\mu) &= \frac{1}{2} \sum_{i,j=1}^k (\langle f_i f_j, \mu \rangle - \langle f_i, \mu \rangle \langle f_j, \mu \rangle) F_{z_i z_j}(\langle \mathbf{f}, \mu \rangle) \\
 &\quad + \sum_{i=1}^k \{\langle Af_i, \mu \rangle + \langle Bf_i, \mu^2 \rangle\} F_{z_i}(\langle \mathbf{f}, \mu \rangle) \\
 &\quad + \sum_{i=1}^k \{\langle (f_i \circ \pi)\sigma, \mu^2 \rangle - \langle f_i, \mu \rangle \langle \sigma, \mu^2 \rangle\} F_{z_i}(\langle \mathbf{f}, \mu \rangle).
 \end{aligned}$$

The formulation (3.14) is from Fleming and Viot (1979), whereas (3.12) is due to Dawson and Hochberg (1982).

Another choice for the domain of \mathcal{L} that is often useful is the set of all $\varphi \in B(\mathcal{P}(E))$ of the form $\varphi(\mu) = \langle f, \mu^k \rangle$, where $k \geq 1$ and $f \in B(E^k)$ satisfies certain conditions. To describe these conditions precisely, we need to be more specific about our assumptions on A . We assume that E is locally compact and that the closure of A generates a Feller semigroup $\{T(t)\}$ on $\hat{C}(E)$, the space of real continuous functions on E vanishing at infinity. (If E is compact, then $\hat{C}(E) = C(E)$.) Note that $\{T(t)\}$ is given by a transition function $P(t, x, d\xi)$, that is,

$$(3.15) \quad T(t)f(x) = \int_E f(\xi) P(t, x, d\xi).$$

For each $k \geq 1$, we define the semigroup $\{T_k(t)\}$ on $B(E^k)$ by

$$(3.16) \quad T_k(t)f(x_1, \dots, x_k) = \int_E \cdots \int_E f(\xi_1, \dots, \xi_k) P(t, x_1, d\xi_1) \cdots P(t, x_k, d\xi_k),$$

and we let $A^{(k)}$ denote its generator; note that $\mathcal{D}(A^{(k)})$ is a subspace of $B(E^k)$.

In addition, for each $k \geq 2$ and $1 \leq i < j \leq k$, we define $\Phi_{ij}^{(k)} : B(E^k) \mapsto B(E^{k-1})$ by letting $\Phi_{ij}^{(k)} f$ be the function obtained from f by replacing x_j by x_i and renumbering the variables:

$$(3.17) \quad (\Phi_{ij}^{(k)} f)(x_1, \dots, x_{k-1}) = f(x_1, \dots, x_{j-1}, x_i, x_j, \dots, x_{k-1}).$$

For each $k \geq 1$ and $1 \leq i \leq k$, we define $H_i^{(k)} : B(E^k) \mapsto B(E^{k+1})$ by

$$(3.18) \quad (H_i^{(k)} f)(x_1, \dots, x_{k+1}) = \int_E f(x_1, \dots, x_{i-1}, \xi, x_{i+1}, \dots, x_k) R((x_i, x_{k+1}), d\xi)$$

and $K_i^{(k)} : B(E^k) \mapsto B(E^{k+2})$ by

$$(3.19) \quad (K_i^{(k)} f)(x_1, \dots, x_{k+2}) = \frac{\bar{\sigma} + \sigma(x_i, x_{k+1}) - \sigma(x_{k+1}, x_{k+2})}{2\bar{\sigma}} f(x_1, \dots, x_k),$$

where $\bar{\sigma} = \sup_{x, y, z \in E} |\sigma(x, y) - \sigma(y, z)|$ and $0/0 = 0$.

For each $k \geq 1$ and $f \in \mathcal{D}(A^{(k)})$, we define $\varphi_f \in B(\mathcal{P}(E))$ by

$$(3.20) \quad \varphi_f(\mu) = \langle f, \mu^k \rangle,$$

and we note that (3.12) reduces (informally, at least) to

$$(3.21) \quad (\mathcal{L}\varphi_f)(\mu) = \sum_{1 \leq i < j \leq k} (\langle \Phi_{ij}^{(k)} f, \mu^{k-1} \rangle - \langle f, \mu^k \rangle) \\ + \langle A^{(k)} f, \mu^k \rangle + \alpha \sum_{i=1}^k (\langle H_i^{(k)} f, \mu^{k+1} \rangle - \langle f, \mu^k \rangle) \\ + 2\bar{\sigma} \sum_{i=1}^k (\langle K_i^{(k)} f, \mu^{k+2} \rangle - \langle f, \mu^k \rangle) + \bar{\sigma} k \langle f, \mu^k \rangle.$$

This leads immediately to the so-called *dual process*. For each $k \geq 1$, $f \in B(E^k)$, and $\mu \in \mathcal{P}(E)$, write

$$(3.22) \quad \varphi_f(\mu) = \varphi_\mu(f) = \langle f, \mu^k \rangle,$$

and consider the Markov process in $\cup_{k=1}^\infty B(E^k)$ with generator $\mathcal{L}^\#$ satisfying

$$(3.23) \quad (\mathcal{L}\varphi_f)(\mu) = (\mathcal{L}^\#\varphi_\mu)(f) + \bar{\sigma}k\varphi_\mu(f).$$

By (3.21), this is a process that jumps from $f \in B(E^k)$ to $\Phi_{ij}^{(k)} f \in B(E^{k-1})$ at rate 1 ($1 \leq i < j \leq k$), to $H_i^{(k)} f \in B(E^{k+1})$ at rate α ($1 \leq i \leq k$), and to $K_i^{(k)} f \in B(E^{k+2})$ at rate $2\bar{\sigma}$ ($1 \leq i \leq k$). Between jumps it moves from $f \in B(E^k)$ to $T_k(t)f \in B(E^k)$ in time t . The process was first described by Dawson and Hochberg (1982).

Let us be more explicit about the dual process. Let $M = \{M(t), t \geq 0\}$ be the pure-jump Markov process in \mathbf{N} , the set of positive integers, with transition intensities $q_{k,k-1} = k(k-1)/2$, $q_{k,k+1} = \alpha k$, and $q_{k,k+2} = 2\bar{\sigma}k$. Let $0 = \tau_0 < \tau_1 < \tau_2 < \dots$ be the jump times of M (this sequence will be finite if $\alpha = \bar{\sigma} = 0$), and given M , let $\Gamma_1, \Gamma_2, \dots$ be a sequence of conditionally independent random operators such that, if $n \geq 1$ and $M(\tau_{n-1}) = k$, then

$$(3.24) \quad \mathbf{P}\{\Gamma_n = \Phi_{ij}^{(k)} \mid M\} = \binom{k}{2}^{-1} \quad (1 \leq i < j \leq k) \quad \text{if } M(\tau_n) = k-1,$$

$$(3.25) \quad \mathbf{P}\{\Gamma_n = H_i^{(k)} \mid M\} = k^{-1} \quad (1 \leq i \leq k) \quad \text{if } M(\tau_n) = k+1,$$

$$(3.26) \quad \mathbf{P}\{\Gamma_n = K_i^{(k)} \mid M\} = k^{-1} \quad (1 \leq i \leq k) \quad \text{if } M(\tau_n) = k+2.$$

The dual process can then be written as

$$(3.27) \quad Y(t) = T_{M(\tau_k)}(t - \tau_k)\Gamma_k T_{M(\tau_{k-1})}(\tau_k - \tau_{k-1})\Gamma_{k-1} \dots \\ \Gamma_2 T_{M(\tau_1)}(\tau_2 - \tau_1)\Gamma_1 T_{M(0)}(\tau_1)Y(0) \quad \text{if } \tau_k \leq t < \tau_{k+1}, \quad k \geq 0,$$

where $Y(0) \in B(E^{M(0)})$. The next result is essentially from Ethier and Kurtz (1987).

THEOREM 3.1. *Let E be locally compact and suppose that the closure of A generates a Feller semigroup on $\hat{C}(E)$. Define B in terms of $\alpha \geq 0$ and a transition function $R((x, y), dx')$ on $E^2 \times \mathcal{B}(E)$ by (3.5), and let $\sigma \in B_{\text{sym}}(E^2)$. Then the martingale problems for \mathcal{L} defined by (3.3) and (3.10), by (3.13) and (3.14), and by (3.20) and (3.21) are equivalent. If $\{\mu_t, t \geq 0\}$ is a solution and if $\{Y(t), t \geq 0\}$ is defined as above and is independent of $\{\mu_t, t \geq 0\}$, then*

$$(3.28) \quad \mathbf{E}[\langle f, \mu_t^m \rangle] = \mathbf{E}\left[\langle Y(t), \mu_0^{M(t)} \rangle \exp\left\{\bar{\sigma} \int_0^t M(s) ds\right\}\right],$$

for all $m \geq 1$, $f \in B(E^m)$, and $t \geq 0$, where $Y(0) = f$ and $M(0) = m$.

A number of authors (Fleming and Viot (1979), Kurtz (1981), Ethier and Kurtz (1987), Donnelly and Kurtz (1993)) have given conditions under which the $C_{\mathcal{P}(E)}[0, \infty)$ martingale problem for \mathcal{L} is well posed, thereby characterizing the diffusion process associated with \mathcal{L} . Because of Theorem 3.1, uniqueness of solutions (with specified initial distribution) is almost immediate. Existence, however, requires a relative compactness argument. The following conditions are essentially from Ethier and Kurtz (1987).

THEOREM 3.2. (a) *Under the assumptions of Theorem 3.1, uniqueness of solutions of the martingale problem for \mathcal{L} defined by (3.13) and (3.14) (with specified initial distribution) holds.*

(b) *Suppose in addition that either $B = 0$ or both E is compact and B maps $C(E)$ into $C(E^2)$. Then the $C_{\mathcal{P}(E)}[0, \infty)$ martingale problem for \mathcal{L} is well posed.*

It is occasionally necessary to simultaneously generalize (3.14) and (3.21). For this purpose we consider the set of all $\varphi \in B(\mathcal{P}(E))$ of the form

$$(3.29) \quad \varphi(\mu) = F(\langle f_1, \mu^{n_1} \rangle, \dots, \langle f_k, \mu^{n_k} \rangle),$$

where $k \geq 1$, $n_1, \dots, n_k \geq 1$, $f_1 \in \mathcal{D}(A^{(n_1)})$, \dots , $f_k \in \mathcal{D}(A^{(n_k)})$, and $F \in C^2(\mathbf{R}^k)$. In this case (3.12) becomes (again, informally)

$$(3.30) \quad \begin{aligned} (\mathcal{L}\varphi)(\mu) &= \frac{1}{2} \sum_{i,j=1}^k \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} (\langle \Psi_{lm}^{(n_i, n_j)}(f_i, f_j), \mu^{n_i+n_j-1} \rangle - \langle f_i, \mu^{n_i} \rangle \langle f_j, \mu^{n_j} \rangle) F_{z_i z_j} \\ &\quad + \sum_{i=1}^k (\mathcal{L}\varphi_{f_i})(\mu) F_{z_i}, \end{aligned}$$

where $\Psi_{lm}^{(n_i, n_j)}(f_i, f_j)$ is the function in $B(E^{n_i+n_j-1})$ obtained from $f_i(x)f_j(y)$ by replacing y_m by x_l and renumbering the variables, $\mathcal{L}\varphi_f$ is as in (3.21), and the partial derivatives of F have the same arguments as F in (3.29). Under the assumptions of Theorem 3.1, the martingale problem for \mathcal{L} defined by (3.29) and (3.30) is equivalent to those of the theorem.

Let $\{\mu_t, t \geq 0\}$ denote the canonical coordinate process on $\Omega = C_{\mathcal{P}(E)}[0, \infty)$. Let \mathcal{M} denote the Borel σ -field, and put $\mathcal{M}_t = \sigma\{\mu_s : 0 \leq s \leq t\}$ for each $t \geq 0$. We now state a result showing that the transformation from the neutral model (no selection) to the selective model is a Girsanov-type transformation. The result is essentially from Dawson (1978).

THEOREM 3.3. *Suppose, in addition to the assumptions of Theorem 3.1, that $\sigma \in B_{\text{sym}}(E^2) \cap \mathcal{D}(A^{(2)})$. Let $P \in \mathcal{P}(\Omega)$ be a solution of the Ω martingale problem for \mathcal{L}_0 defined by (3.20) and (3.21) with $\bar{\sigma} = 0$. Then*

$$(3.31) \quad R_t = \exp \left\{ \frac{1}{2} \varphi_\sigma(\mu_t) - \frac{1}{2} \varphi_\sigma(\mu_0) - \int_0^t (e^{-\varphi_\sigma/2} \mathcal{L}_0 e^{\varphi_\sigma/2})(\mu_s) ds \right\}$$

is a mean-one $\{\mathcal{M}_t\}$ -martingale on (Ω, \mathcal{M}, P) . Moreover, if we define $Q \in \mathcal{P}(\Omega)$ by $dQ/dP|_{\mathcal{M}_t} = R_t$ for all $t \geq 0$, then Q is a solution of the Ω martingale problem for \mathcal{L} defined by (3.20) and (3.21).

The first assertion follows from Lemma 4.3.2 of Ethier and Kurtz (1986). The fact that Q corresponds to \mathcal{L} is essentially a consequence of the calculation

$$(3.32) \quad \begin{aligned} &\frac{\mathcal{L}_0(\varphi_f e^{\varphi_\sigma/2}) - \varphi_f \mathcal{L}_0 e^{\varphi_\sigma/2}}{e^{\varphi_\sigma/2}} \\ &= \mathcal{L}_0 \varphi_f + \frac{\mathcal{L}_0(\varphi_f e^{\varphi_\sigma/2}) - \varphi_f \mathcal{L}_0 e^{\varphi_\sigma/2} - e^{\varphi_\sigma/2} \mathcal{L}_0 \varphi_f}{e^{\varphi_\sigma/2}} = \mathcal{L} \varphi_f, \end{aligned}$$

where the last step uses (3.30). Also by (3.30), the integrand in (3.31) reduces to

$$(3.33) \quad \begin{aligned} &(e^{-\varphi_\sigma/2} \mathcal{L}_0 e^{\varphi_\sigma/2})(\mu) \\ &= \frac{1}{2} (\langle \Phi_{12}^{(2)} \sigma, \mu \rangle + \langle A^{(2)} \sigma, \mu^2 \rangle + 2\alpha \langle H_1^{(2)} \sigma, \mu^3 \rangle + 2\bar{\sigma} \langle K_1^{(2)} \sigma, \mu^4 \rangle \\ &\quad - (1 + 2\alpha + \bar{\sigma}) \langle \sigma, \mu^2 \rangle). \end{aligned}$$

We consider separately the special case in which $B = 0$ and $\sigma \equiv 0$. The following result seems to be new.

THEOREM 3.4. *Let E be compact, and suppose that the closure of A generates a Feller semigroup on $C(E)$. Then the closure of the operator \mathcal{L} defined on*

$$(3.34) \quad \mathcal{D}(\mathcal{L}) = \{\varphi_f : f \in \mathcal{D}(A^{(k)}) \cap C(E^k), k \geq 1\}$$

by (3.21) with $\alpha = \bar{\sigma} = 0$ generates a Feller semigroup on $C(\mathcal{P}(E))$.

Proof. We apply the Hille–Yosida theorem. $\mathcal{D}(\mathcal{L})$ is dense in $C(\mathcal{P}(E))$ since $\mathcal{D}(A^{(k)}) \cap C(E^k)$ is dense in $C(E^k)$ for each $k \geq 1$. Dissipativity of \mathcal{L} follows by (3.11) and the fact that the closure of \mathcal{L} defined by (3.3), (3.10), and linearity extends \mathcal{L} defined by (3.34) and (3.21). Let $\lambda > 0$. We need to show that $\mathcal{R}(\lambda - \mathcal{L})$ is dense in $C(\mathcal{P}(E))$. Given $k \geq 1$ and $g \in C(E^k)$, define $f \in \mathcal{D}(A^{(k)}) \cap C(E^k)$ by

$$(3.35) \quad f = \sum_{n=0}^{\infty} \left[(\lambda + \binom{k}{2} - A^{(k)})^{-1} \sum_{1 \leq i < j \leq k} \Psi_{ij}^{(k)} \right]^n (\lambda + \binom{k}{2} - A^{(k)})^{-1} g,$$

where $\Psi_{ij}^{(k)} : C(E^k) \mapsto C(E^k)$ is defined by letting $\Psi_{ij}^{(k)} f$ be the function obtained from f by replacing x_j by x_i . Then

$$(3.36) \quad (\lambda + \binom{k}{2} - A^{(k)})f - \sum_{1 \leq i < j \leq k} \Psi_{ij}^{(k)} f = g,$$

so $(\lambda - \mathcal{L})\varphi_f = \varphi_g$, and the range condition follows. Thus, the closure of \mathcal{L} generates a strongly continuous conservative contraction semigroup $\{\mathcal{T}(t)\}$ on $C(\mathcal{P}(E))$. If $k \geq 1$, $f \in C(E^k)$, and $\varphi_f \geq 0$, then

$$(3.37) \quad \|\varphi_f\| - \mathcal{T}(t)\varphi_f(\mu) = \mathcal{T}(t)(\|\varphi_f\| - \varphi_f)(\mu) \leq \|\|\varphi_f\| - \varphi_f\| \leq \|\varphi_f\|$$

for all $\mu \in \mathcal{P}(E)$ and $t \geq 0$, so the positivity of $\{\mathcal{T}(t)\}$ follows. (Note that $\varphi_f \geq 0$ does not imply $f \geq 0$.) \square

4. Convergence. We begin by showing that, under suitable conditions, the Fleming–Viot process of §3 approximates the Markov chains of §2.

LEMMA 4.1. *For each $M \geq 1$, let $P_M(\mu, d\nu)$ be a one-step transition function on $\mathcal{P}_M(E) \times \mathcal{B}(\mathcal{P}_M(E))$ (see (2.2.2)) and let $\beta_M > 0$. Assume that $\lim_{M \rightarrow \infty} \beta_M = \infty$. Let $\mathcal{D} \subset B(E)$ and suppose that for each $f, g \in \mathcal{D}$ there exist $a_{f,g}, b_f \in B(\mathcal{P}(E))$ such that*

$$(4.1) \quad \beta_M \int (\langle f, \nu \rangle - \langle f, \mu \rangle) P_M(\mu, d\nu) = b_f(\mu) + o(1),$$

$$(4.2) \quad \beta_M \int (\langle f, \nu \rangle - \langle f, \mu \rangle)(\langle g, \nu \rangle - \langle g, \mu \rangle) P_M(\mu, d\nu) = a_{f,g}(\mu) + o(1),$$

$$(4.3) \quad \beta_M \int (\langle f, \nu \rangle - \langle f, \mu \rangle)^4 P_M(\mu, d\nu) = o(1),$$

as $M \rightarrow \infty$, uniformly in $\mu \in \mathcal{P}_M(E)$. Let $\varphi \in B(\mathcal{P}(E))$ have the form

$$(4.4) \quad \varphi(\mu) = F(\langle f_1, \mu \rangle, \dots, \langle f_k, \mu \rangle) = F(\langle \mathbf{f}, \mu \rangle),$$

where $k \geq 1$, $f_1, \dots, f_k \in \mathcal{D}$, and $F \in C^2(\mathbf{R}^k)$, and define

$$(4.5) \quad (\mathcal{L}_M \varphi)(\mu) = \beta_M \int (\varphi(\nu) - \varphi(\mu)) P_M(\mu, d\nu)$$

and

$$(4.6) \quad (\mathcal{L}_\infty \varphi)(\mu) = \frac{1}{2} \sum_{i,j=1}^k a_{f_i, f_j}(\mu) F_{z_i z_j}(\langle \mathbf{f}, \mu \rangle) + \sum_{i=1}^k b_{f_i}(\mu) F_{z_i}(\langle \mathbf{f}, \mu \rangle).$$

Then

$$(4.7) \quad (\mathcal{L}_M \varphi)(\mu) = (\mathcal{L}_\infty \varphi)(\mu) + o(1)$$

as $M \rightarrow \infty$, uniformly in $\mu \in \mathcal{P}_M(E)$.

Proof. The result is immediate from a second-order Taylor expansion. \square

Suppose that the assumptions of Theorem 3.2(b) hold, that the sequences $\{w_M\}$, $\{R_M\}$, and $\{Q_M\}$ satisfy (3.6)–(3.8), and that σ is continuous. Suppose also that $\{\mu_t, t \geq 0\}$ is a solution of the $C_{\mathcal{P}(E)}[0, \infty)$ martingale problem for the operator \mathcal{L}_∞ defined by (4.4) and (4.6) with $\mathcal{D} = \mathcal{D}(A)$,

$$(4.8) \quad a_{f,g}(\mu) = \langle fg, \mu \rangle - \langle f, \mu \rangle \langle g, \mu \rangle,$$

and

$$(4.9) \quad b_f(\mu) = c\{\langle Af, \mu \rangle + \langle Bf, \mu^2 \rangle + \langle (f \circ \pi)\sigma, \mu^2 \rangle - \langle f, \mu \rangle \langle \sigma, \mu^2 \rangle\},$$

where c is a positive constant.

Let $\{\nu_\tau^{(N)}, \tau \in \mathbf{Z}_+\}$ ($N = 1, 2, \dots$) be a sequence of diploid models as described in §2.1, and suppose $c = 1$. If $\hat{\nu}_0^{(N)} \pi^{-1} \Rightarrow \mu_0$, then

$$(4.10) \quad \{\hat{\nu}_{[2Nt]}^{(N)} \pi^{-1}, t \geq 0\} \Rightarrow \{\mu_t, t \geq 0\}$$

in $D_{\mathcal{P}(E)}[0, \infty)$. This follows from a slight modification of Lemma 4.1 (or of (3.4)), together with Corollary 4.8.17 of Ethier and Kurtz (1986).

Let $\{\nu_\tau^{(M)}, \tau \in \mathbf{Z}_+\}$ ($M = 1, 2, \dots$) be a sequence of Wright–Fisher models as described in §2.2, and suppose $c = 1$. If $\nu_0^{(M)} \Rightarrow \mu_0$, then

$$(4.11) \quad \{\nu_{[Mt]}^{(M)}, t \geq 0\} \Rightarrow \{\mu_t, t \geq 0\}$$

in $D_{\mathcal{P}(E)}[0, \infty)$. This follows from Lemma 4.1 or (3.11), together with Corollary 4.8.17 of Ethier and Kurtz (1986).

Let $\{\nu_\tau^{(M)}, \tau \in \mathbf{Z}_+\}$ ($M = 1, 2, \dots$) be a sequence of Moran models as described in §2.3, and suppose $c = \frac{1}{2}$. If $\nu_0^{(M)} \Rightarrow \mu_0$, then

$$(4.12) \quad \{\nu_{[M^2 t/2]}^{(M)}, t \geq 0\} \Rightarrow \{\mu_t, t \geq 0\}$$

in $D_{\mathcal{P}(E)}[0, \infty)$. This follows from Lemma 4.1, together with Corollary 4.8.17 of Ethier and Kurtz (1986).

There are two problems with the above diffusion approximations. First, the topology of weak convergence is too weak for many applications. For example, we might be interested in functions φ on $\mathcal{P}(E)$ such as

$$(4.13) \quad \varphi(\mu) = \mu^2(D), \quad D = \{(x, y) \in E^2 : x = y\},$$

the population homozygosity. But this function is not continuous in the topology of weak convergence (unless D is open, as, e.g., when E is finite). Second, the assumption that the selection intensity function σ is continuous rules out a number of important examples, e.g.,

$$(4.14) \quad \sigma(x, y) = \sigma I_D(x, y),$$

where σ is a real constant and D is as in (4.13); this corresponds to the situation in which homozygotes have a selective advantage or disadvantage compared to heterozygotes. We treat these two issues separately.

A solution to the first problem is obtained by imposing a stronger topology on $\mathcal{P}(E)$. Here we assume that (E, r) is a complete separable metric space. Let ρ denote the Prohorov metric on $\mathcal{P}(E)$, which induces the topology of weak convergence. Define the metric ρ_a on $\mathcal{P}(E)$ by

$$(4.15) \quad \rho_a(\mu, \nu) = \rho(\mu, \nu) + \sup_{0 < \varepsilon \leq 1} \left| \int_E \int_E \left(1 - \frac{r(x, y)}{\varepsilon}\right)_+ \mu(dx) \mu(dy) - \int_E \int_E \left(1 - \frac{r(x, y)}{\varepsilon}\right)_+ \nu(dx) \nu(dy) \right|.$$

The topology on $\mathcal{P}(E)$ induced by ρ_a is called the *weak atomic topology* (Ethier and Kurtz (1993a)).

LEMMA 4.2. *Let $\{\mu_n\} \subset \mathcal{P}(E)$ and $\mu \in \mathcal{P}(E)$. The following are equivalent:*

(a) $\rho_a(\mu_n, \mu) \rightarrow 0$.

(b) $\rho(\mu_n, \mu) \rightarrow 0$ and $\sum_x \mu_n(\{x\})^2 \rightarrow \sum_x \mu(\{x\})^2$.

(c) $\rho(\mu_n, \mu) \rightarrow 0$ and the sizes and locations of the atoms of μ_n converge to the sizes and locations of the atoms of μ .

For example, if $x_n \rightarrow x$, $x_n \neq x$ for all n , and $\mu_n = \frac{1}{2}(\delta_{x_n} + \delta_x)$, then $\rho(\mu_n, \delta_x) \rightarrow 0$ but $\rho_a(\mu_n, \delta_x) \not\rightarrow 0$.

Let $k \geq 1$, $\mathbf{n} = (n_1, \dots, n_k) \in \mathbf{N}^k$, and $n = n_1 + \dots + n_k > 1$, and define $\Gamma_{\mathbf{n}} = \{(x_1, \dots, x_n) \in E^n : \text{there exist distinct } y_1, \dots, y_k \in E \text{ such that } |\{1 \leq i \leq n : x_i = y_j\}| = n_j \text{ for } j = 1, \dots, k\}$. Then the mapping $\varphi_{\mathbf{n}} : \mathcal{P}(E) \mapsto [0, 1]$ given by $\varphi_{\mathbf{n}}(\mu) = \mu^n(\Gamma_{\mathbf{n}})$ is continuous in the weak atomic topology but not typically in the topology of weak convergence. Of course, (4.13) is a special case. See §9.2 for the relevance of such functions.

The result we need in order to generalize the diffusion approximations discussed above is the following theorem.

THEOREM 4.3. *A sequence of processes $\{\mu_t^{(M)}, t \geq 0\}$ ($M = 1, 2, \dots$) with sample paths in $D_{(\mathcal{P}(E), \rho_a)}[0, \infty)$ is relatively compact in $D_{(\mathcal{P}(E), \rho_a)}[0, \infty)$ if and only if it is relatively compact in $D_{(\mathcal{P}(E), \rho)}[0, \infty)$ and for each $T > 0$ and $\delta > 0$ there exists $\varepsilon > 0$ such that*

$$(4.16) \quad \liminf_{M \rightarrow \infty} \mathbf{P}\left\{ \sup_{0 \leq t \leq T} (\mu_t^{(M)} \times \mu_t^{(M)})(\{(x, y) \in E^2 : 0 < r(x, y) < \varepsilon\}) \leq \delta \right\} \geq 1 - \delta.$$

To establish relative compactness in $D_{(\mathcal{P}(E), \rho_a)}[0, \infty)$ for the three models of §2, additional assumptions are needed. For each $M \geq 1$, write

$$(4.17) \quad Q_M(x, \cdot) = \left(1 - \frac{\theta_M(x)}{2M}\right) \delta_x(\cdot) + \frac{\theta_M(x)}{2M} \lambda_M(x, \cdot),$$

where $\theta_M \in B(E)$ is nonnegative and λ_M is a one-step transition function on $E \times \mathcal{B}(E)$, and suppose that

$$(4.18) \quad \sup_M \sup_{x \in E} \theta_M(x) < \infty,$$

$$(4.19) \quad \lim_{\varepsilon \rightarrow 0} \limsup_{M \rightarrow \infty} \sup_{x, y \in E} \theta_M(x) \lambda_M(x, B_\varepsilon(y)) = 0,$$

and

$$(4.20) \quad \lim_{\varepsilon \rightarrow 0} \sup_{x, y, z \in E} R((x, y), B_\varepsilon(z)) = 0,$$

where $B_\varepsilon(y)$ is the ball of radius ε centered at y . These assumptions, in addition to the ones imposed previously, guarantee that (4.10)–(4.12) hold in $D_{(\mathcal{P}(E), \rho_a)}[0, \infty)$; see Ethier and Kurtz (1993a).

We turn next to the problem of relaxing the continuity assumption on the selection intensity function σ . Let E be compact, let $\theta > 0$ and $\nu_0 \in \mathcal{P}(E)$, and put

$$(4.21) \quad (Af)(x) = \frac{1}{2} \theta \int_E (f(\xi) - f(x)) \nu_0(d\xi).$$

We assume all previously stated assumptions in this section except for the continuity of σ , and we assume in addition that A is given by (4.21) and $B = 0$. Define $\sigma_0 \in B(E)$ by $\sigma_0(x) = \sigma(x, x)$, and let $D_{\sigma_0} = \{x \in E : \sigma_0 \text{ is discontinuous at } x\}$ and $D_{\sigma}^- = \{(x, y) \in E^2 : \sigma \text{ is discontinuous at } (x, y) \text{ and } x \neq y\}$. Assume further that

$$(4.22) \quad \mathbf{P}\{(\mu_0 + \nu_0)(D_{\sigma_0}) = 0\} = 1$$

and

$$(4.23) \quad \mathbf{P}\{(\mu_0 + \nu_0)^2(D_{\sigma}^-) = 0\} = 1.$$

Then (4.10)–(4.12) hold in $D_{(\mathcal{P}(E), \rho_a)}[0, \infty)$.

As for the proof, relative compactness in $D_{(\mathcal{P}(E), \rho_a)}[0, \infty)$ is established as before. The difficulty is in identifying the limiting process. The assumptions (4.22) and (4.23) ensure that $\mathcal{L}\varphi_f$ is continuous in the weak atomic topology almost surely with respect to the distribution of μ_t for each $f \in C(E^k)$, $k \geq 1$, and $t \geq 0$. This is proved using Theorem 3.3 above and Theorem 8.3 below, the latter explaining why (4.21) is assumed. But undoubtedly the conclusions hold in greater generality.

Finally, convergence in distribution of sequences of Fleming–Viot processes as well as weak convergence of stationary distributions can be established under similar conditions to the above. See Ethier and Kurtz (1993a). One technique we have not discussed here is the use of dual processes to prove convergence in distribution of

sequences of Fleming-Viot processes. In particular, this is well adapted to the case of discontinuous selection intensity functions. See Ethier and Kurtz (1987).

5. Ergodicity. The ergodic theorem for a measurable E -valued stationary process Z states that if f is a real-valued Borel function on E with $\mathbf{E}[|f(Z(0))|] < \infty$, then

$$(5.1) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t f(Z(s)) ds = \mathbf{E}[f(Z(0)) \mid \mathcal{I}],$$

where \mathcal{I} is the σ -algebra of invariant events, that is, events of the form $\{Z(\cdot) \in \Gamma\}$, $\Gamma \in \mathcal{B}(E^{[0, \infty)})$, such that $\{Z(\cdot) \in \Gamma\} = \{Z(t + \cdot) \in \Gamma\}$ for all $t \geq 0$. If the events in \mathcal{I} are all of probability zero or one, then Z is said to be *ergodic* and the right-hand side of (5.1) is just $\mathbf{E}[f(Z(0))]$. (See, e.g., Durrett (1991, Chap. 6).)

If $P(t, x, d\xi)$ is the transition function for an E -valued Markov process and $\nu_0 \in \mathcal{P}(E)$ is a stationary distribution for $P(t, x, d\xi)$, that is,

$$(5.2) \quad \int_E P(t, x, \Gamma) \nu_0(dx) = \nu_0(\Gamma), \quad \Gamma \in \mathcal{B}(E),$$

then a sufficient condition for ergodicity of the stationary process obtained by taking ν_0 to be the initial distribution is that ν_0 be the unique probability measure satisfying (5.2). If the closure of A generates the semigroup corresponding to $P(t, x, d\xi)$ on a subspace of $B(E)$ that is separating, then (5.2) is equivalent to

$$(5.3) \quad \langle Af, \nu_0 \rangle = 0, \quad f \in \mathcal{D}(A).$$

The following is essentially a result of Dynkin (1989).

THEOREM 5.1. *Let E be compact, and suppose that the closure of A generates a Feller semigroup on $C(E)$. Suppose further that there exists a unique $\nu_0 \in \mathcal{P}(E)$ satisfying (5.3). Then there exists a unique stationary distribution for the Fleming-Viot process with type space E and mutation operator A (and no recombination or selection).*

Proof. Existence is automatic by virtue of the Feller property of the Fleming-Viot process. For uniqueness, we must show that there is a unique $\Pi \in \mathcal{P}(\mathcal{P}(E))$ satisfying

$$(5.4) \quad \langle \mathcal{L}\varphi_f, \Pi \rangle = 0$$

for all $k \geq 1$ and $f \in \mathcal{D}(A^{(k)}) \cap C(E^k)$, where $\mathcal{L}\varphi_f$ is as in (3.21) with $\alpha = \bar{\sigma} = 0$. For $k = 1$, (5.4) becomes

$$(5.5) \quad \int_{\mathcal{P}(E)} \langle Af, \mu \rangle \Pi(d\mu) = 0, \quad f \in \mathcal{D}(A),$$

which by the assumption on ν_0 implies that $\int_{\mathcal{P}(E)} \mu(\Gamma) \Pi(d\mu) = \nu_0(\Gamma)$ for all $\Gamma \in \mathcal{B}(E)$ and hence that

$$(5.6) \quad \int_{\mathcal{P}(E)} \langle f, \mu \rangle \Pi(d\mu) = \langle f, \nu_0 \rangle, \quad f \in C(E).$$

We proceed by induction. Given $k \geq 2$, suppose that $\int_{\mathcal{P}(E)} \langle f, \mu^{k-1} \rangle \Pi(d\mu)$ is determined for all $f \in C(E^{k-1})$. By (5.4),

$$(5.7) \quad \int_{\mathcal{P}(E)} \langle \binom{k}{2} f - A^{(k)} f, \mu^k \rangle \Pi(d\mu) = \int_{\mathcal{P}(E)} \sum_{1 \leq i < j \leq k} \langle \Phi_{ij}^{(k)} f, \mu^{k-1} \rangle \Pi(d\mu)$$

for all $f \in \mathcal{D}(A^{(k)}) \cap C(E^k)$, but since the range of $\lambda - A^{(k)}$ contains $C(E^k)$ for all $\lambda > 0$, (5.7) implies that

$$(5.8) \quad \int_{\mathcal{P}(E)} \langle f, \mu^k \rangle \Pi(d\mu) = \int_{\mathcal{P}(E)} \sum_{1 \leq i < j \leq k} \langle \Phi_{ij}^{(k)} R^{(k)} f, \mu^{k-1} \rangle \Pi(d\mu)$$

for all $f \in C(E^k)$, where

$$(5.9) \quad R^{(k)} f = (\binom{k}{2} - A^{(k)})^{-1} f = \int_0^\infty e^{-(\binom{k}{2})t} T_k(t) f dt.$$

Uniqueness of stationary distributions follows. \square

For a Markov process Z with transition function $P(t, x, d\xi)$ we are typically interested in statements stronger than just uniqueness of stationary distributions. In particular, we would like to know that for an arbitrary initial distribution, the process is, in some sense, asymptotically stationary. Two possible senses are

$$(5.10) \quad \lim_{t \rightarrow \infty} \mathbf{E}[f(Z(t))] = \langle f, \nu_0 \rangle, \quad f \in \overline{C}(E),$$

for every initial distribution, which we refer to as *weak ergodicity* (occasionally we may need to replace the limit in (5.10) by the corresponding Cesàro limit), and

$$(5.11) \quad \lim_{t \rightarrow \infty} \|P(t, x, \cdot) - \nu_0\|_{\text{var}} = 0, \quad x \in E,$$

which we refer to as *strong ergodicity*. Here $\|\nu_1 - \nu_2\|_{\text{var}} = \sup_{\Gamma \in \mathcal{B}(E)} |\nu_1(\Gamma) - \nu_2(\Gamma)|$. One important consequence of strong ergodicity is that

$$(5.12) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t f(Z(s)) ds = \langle f, \nu_0 \rangle \quad \text{a.s.}, \quad f \in B(E),$$

for every initial distribution.

Duality gives a simple proof of weak ergodicity in the absence of recombination and selection.

THEOREM 5.2. *Let E be compact, and suppose that the closure of A generates a Feller semigroup $\{T(t)\}$ on $C(E)$ corresponding to a weakly ergodic Markov process (equivalently, there exists $\nu_0 \in \mathcal{P}(E)$ such that $\lim_{t \rightarrow \infty} T(t)f(x) = \langle f, \nu_0 \rangle$ for all $f \in C(E)$ and $x \in E$). Then the Fleming–Viot process with type space E and mutation operator A is weakly ergodic.*

Proof. Note that without recombination or selection, the Markov jump process M in the definition of the dual process is a pure death process absorbing at 1. Let τ be the absorption time. Then by the duality identity (3.28) and the assumption of weak ergodicity, for each $m \geq 1$ and $f \in C(E^m)$ we have

$$(5.13) \quad \lim_{t \rightarrow \infty} \mathbf{E}[\langle f, \mu_t^m \rangle] = \lim_{t \rightarrow \infty} \mathbf{E}[\langle T(t - \tau)Y(\tau), \mu_0 \rangle; \tau \leq t] = \mathbf{E}[\langle Y(\tau), \nu_0 \rangle; \tau < \infty],$$

regardless of the distribution of μ_0 . Together with the Riesz representation theorem, this proves the theorem. \square

Coupling arguments provide one approach to strong ergodicity. The basic idea is to construct E -valued Markov processes Z_1 and Z_2 with transition function $P(t, x, d\xi)$ but with different initial distributions on the same probability space in such a way that there is a random time τ such that $Z_1(t) = Z_2(t)$ for all $t \geq \tau$. We then have the basic coupling inequality

$$(5.14) \quad \|\mathbf{P}\{Z_1(t) \in \cdot\} - \mathbf{P}\{Z_2(t) \in \cdot\}\|_{\text{var}} \leq \mathbf{P}\{\tau > t\}, \quad t \geq 0.$$

If $Z_1(0)$ has distribution δ_x and $Z_2(0)$ has distribution ν_0 (a stationary distribution), then

$$(5.15) \quad \|P(t, x, \cdot) - \nu_0\|_{\text{var}} \leq \mathbf{P}\{\tau > t\}, \quad t \geq 0.$$

We say that the coupling is *successful* if $\mathbf{P}\{\tau < \infty\} = 1$. We call the coupling a *Markov coupling* if the coupled pair is a Markov process in $E \times E$. See Griffeath (1978) for a general discussion of coupling. The following theorem is from Ethier and Kurtz (1993b).

THEOREM 5.3. *Let E be locally compact, and suppose that the closure of A generates a Feller semigroup on $\hat{C}(E)$. Let \tilde{A} be the generator for a Markov process in $E \times E$ giving a Markov coupling for A , that is, if (Z_1, Z_2) is a Markov process corresponding to \tilde{A} with initial distribution $\nu_1 \times \nu_2$, where $\nu_1, \nu_2 \in \mathcal{P}(E)$, then Z_i is a Markov process corresponding to A with initial distribution ν_i for $i = 1, 2$. Let $\{\tilde{\mu}_t, t \geq 0\}$ be a Fleming–Viot process with type space $E \times E$ and mutation operator \tilde{A} (and no recombination or selection). Then $\mu_t^{(1)}(\Gamma) = \tilde{\mu}_t(\Gamma \times E)$ and $\mu_t^{(2)}(\Gamma) = \tilde{\mu}_t(E \times \Gamma)$ define Fleming–Viot processes with type space E and mutation operator A . If the coupling given by \tilde{A} is successful, then there exists a random time $\tau < \infty$ a.s. such that $\tilde{\mu}_t\{(x, x) : x \in E\} = 1$ for all $t \geq \tau$. In particular, $\mu_t^{(1)} = \mu_t^{(2)}$ for all $t \geq \tau$, so the Fleming–Viot process with type space E and mutation operator A is strongly ergodic.*

The simplest example satisfying the hypotheses of the theorem is

$$(5.16) \quad (Af)(x) = \frac{1}{2}\theta \int_E (f(\xi) - f(x)) \nu_0(d\xi)$$

with

$$(5.17) \quad (\tilde{A}f)(x, y) = \frac{1}{2}\theta \int_E (f(\xi, \xi) - f(x, y)) \nu_0(d\xi),$$

where $\theta > 0$ and $\nu_0 \in \mathcal{P}(E)$. In this case the coupling inequality (5.15) gives the rate of convergence to equilibrium for the Fleming–Viot process; see (8.11) below.

Theorem 5.3 can be extended to models with recombination under a somewhat stronger assumption on the coupling \tilde{A} . See Ethier and Kurtz (1993b).

There does not seem to be a direct extension of Theorem 5.3 to models with selection. In particular, there does not seem to be a selection intensity function on the product space $E \times E$ for which both marginal processes are Fleming–Viot processes in $\mathcal{P}(E)$ with the desired mutation and selection. We can, however, define a selection intensity function on $E \times E$ for which one of the marginal processes has the desired properties. For $\sigma \in B_{\text{sym}}(E \times E)$, define $\tilde{\sigma}_i \in B_{\text{sym}}(E^2 \times E^2)$ for $i = 1, 2$ by $\tilde{\sigma}_i((x_1, x_2), (y_1, y_2)) = \sigma(x_i, y_i)$.

Let the coupled neutral model of Theorem 5.3 be defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and let $\sigma \in B_{\text{sym}}(E \times E) \cap \mathcal{D}(A^{(2)})$. For $i = 1, 2$, define $\tilde{\sigma}_i$ as above, define \tilde{R}_t^i as in (3.31) using $\tilde{\sigma}_i$, and set $d\mathbf{Q}_i = \tilde{R}_t^i d\mathbf{P}$ on the σ -algebra generated by $\{\tilde{\mu}_s : 0 \leq s \leq t\}$. Since on any time interval $[0, t]$ the Radon–Nikodym derivatives are bounded below by a constant $\rho(t)$, we see that

$$(5.18) \quad \mathbf{Q}_i\{\mu_t^{(i)} \in \Gamma\} \geq \rho(t)\mathbf{P}\{\mu_t^{(i)} \in \Gamma\} \geq \rho(t)\mathbf{P}\{\mu_t^{(i)} \in \Gamma, \tau \leq t\}.$$

But the right-hand side of (5.18) does not depend on i , so if t is large enough that $\mathbf{P}\{\tau \leq t\} > 0$, we see that the distributions $\mathbf{Q}_1(\mu_t^{(1)})^{-1}$ and $\mathbf{Q}_2(\mu_t^{(2)})^{-1}$ cannot be mutually singular. It follows that the Fleming–Viot process with mutation operator A and selection intensity function σ can have at most one stationary distribution. If there exist $t > 0$ and $\varepsilon > 0$ such that $\mathbf{P}\{\tau \leq t\} \geq \varepsilon$ for all choices of the distribution of $\tilde{\mu}_0$, then a renewal argument gives the following theorem from Ethier and Kurtz (1993b).

THEOREM 5.4. *Suppose the assumptions of Theorem 5.3 hold. Let $\{\tilde{\mu}_t, t \geq 0\}$ be as in that theorem, and assume that there exist $t > 0$ and $\varepsilon > 0$ such that $\mathbf{P}\{\tau \leq t\} \geq \varepsilon$ for all choices of the initial distribution. Let $\sigma \in B_{\text{sym}}(E \times E) \cap \mathcal{D}(A^{(2)})$. Then the Fleming–Viot process with type space E , mutation operator A , and selection intensity function σ is strongly ergodic.*

6. An infinite particle system. Let E be compact. (Recall that if the desired type space is only locally compact, e.g., \mathbf{R}^d , then E can be taken to be its one-point compactification.) With reference to (3.21), observe that in the case of no recombination or selection,

$$(6.1) \quad (\mathcal{L}\varphi_f)(\mu) = \sum_{1 \leq i < j \leq n} (\langle \Phi_{ij}^{(n)} f, \mu^{n-1} \rangle - \langle f, \mu^n \rangle) + \langle A^{(n)} f, \mu^n \rangle = \langle Cf, \mu^n \rangle,$$

for all $f \in \mathcal{D}(A^{(n)})$, where

$$(6.2) \quad (Cf)(x_1, \dots, x_n) = \sum_{1 \leq i < j \leq n} \{f(\theta_{ij}(x_1, \dots, x_n)) - f(x_1, \dots, x_n)\} + (A^{(n)}f)(x_1, \dots, x_n),$$

$\theta_{ij}(x_1, \dots, x_n)$ being the element of E^n obtained from (x_1, \dots, x_n) by replacing the j th component by the i th.

We interpret C as an operator with domain in $B(E^\infty)$. It is clear that C is the generator for an E^∞ -valued process $X = (X_1, X_2, \dots)$ whose j th component (which we refer to as the particle at level j) evolves according to the mutation process until it “looks down” to level i for some $i < j$ and changes its value to the value on level i . This process was first considered by Dawson and Hochberg (1982), and it played a key role in their work on the support properties of the Fleming–Viot process with Brownian mutation. It was studied in depth in Donnelly and Kurtz (1993), and results from that paper are summarized here.

Let $\{S(t)\}$ denote the Feller semigroup defined on $C(E^\infty)$ corresponding to C . Then, viewing $C(E^n)$ as a closed subspace of $C(E^\infty)$, we note that $S(t) : C(E^n) \mapsto C(E^n)$ for all $t \geq 0$. The fact that (6.1) holds of course implies for each $\lambda > 0$ that

$$(6.3) \quad (\lambda - \mathcal{L})\varphi_f(\mu) = \langle (\lambda - C)f, \mu^n \rangle$$

for all $f \in \mathcal{D}(A^{(n)}) \cap C(E^n)$ and hence that

$$(6.4) \quad (\lambda - \mathcal{L})^{-1} \varphi_f(\mu) = \langle (\lambda - C)^{-1} f, \mu^n \rangle$$

for all $f \in C(E^n)$. This last identity ensures that

$$(6.5) \quad \mathbf{E}_\mu[\langle f, \mu_t^n \rangle] = \langle S(t)f, \mu^n \rangle = \mathbf{E}_{\mu^\infty}^X[f(X_1(t), \dots, X_n(t))]$$

for all $f \in C(E^n)$ and $t \geq 0$. Here \mathbf{E}_μ denotes the expectation for the Fleming–Viot process under the assumption that the initial state is μ , and $\mathbf{E}_{\mu^\infty}^X$ denotes the expectation for the particle system under the assumption that $X_1(0), X_2(0), \dots$ are i.i.d. with common distribution μ . Let $\nu \in \mathcal{P}(\mathcal{P}(E))$. By (6.5),

$$(6.6) \quad \int_{\mathcal{P}(E)} \mathbf{E}_\mu[\langle f, \mu_t^n \rangle] \nu(d\mu) = \int_{\mathcal{P}(E)} \mathbf{E}_{\mu^\infty}^X[f(X_1(t), \dots, X_n(t))] \nu(d\mu)$$

for all $f \in B(E^n)$ and $t \geq 0$. The left-hand side is just the expectation for a Fleming–Viot process with initial distribution ν , and the right-hand side is the expectation for the particle system under the assumption that $(X_1(0), X_2(0), \dots)$ is an exchangeable sequence with

$$(6.7) \quad \mathbf{P}\{X_1(0) \in \Gamma_1, \dots, X_n(0) \in \Gamma_n\} = \int_{\mathcal{P}(E)} \prod_{i=1}^n \mu(\Gamma_i) \nu(d\mu).$$

The identity (6.6) implies that

$$(6.8) \quad \mathbf{P}\{X_1(t) \in \Gamma_1, \dots, X_n(t) \in \Gamma_n\} = \int_{\mathcal{P}(E)} \prod_{i=1}^n \mu(\Gamma_i) \nu_t(d\mu)$$

for all $t \geq 0$, where ν_t is the distribution at time t of the Fleming–Viot process with initial distribution ν . Consequently, we see that if $(X_1(0), X_2(0), \dots)$ is an exchangeable sequence, then $(X_1(t), X_2(t), \dots)$ is an exchangeable sequence, and the corresponding de Finetti measure

$$(6.9) \quad \hat{\mu}_t = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$$

has the same distribution as μ_t . In fact, as the next theorem states, the process $\{\hat{\mu}_t, t \geq 0\}$ is a version of the Fleming–Viot process.

THEOREM 6.1. *Let E be compact, and suppose that the closure of A generates a Feller semigroup on $C(E)$. Let $X = (X_1, X_2, \dots)$ be a Markov process in E^∞ with generator C and suppose that $(X_1(0), X_2(0), \dots)$ is exchangeable. Then, for each $t > 0$, $(X_1(t), X_2(t), \dots)$ is exchangeable, and the process given by the de Finetti measures (6.9) is a Fleming–Viot process with type space E and mutation operator A .*

Proof. Note that the process given by the first n components of X is itself a Markov process. In Donnelly and Kurtz (1993) it is shown that this E^n -valued process can be coupled to an E^n -valued process $Y^{(n)}$ with generator

$$(6.10) \quad \begin{aligned} &(\tilde{C}^n f)(x_1, \dots, x_n) \\ &= \frac{1}{2} \sum_{i \neq j} \{f(\theta_{ij}(x_1, \dots, x_n)) - f(x_1, \dots, x_n)\} + (A^{(n)} f)(x_1, \dots, x_n) \end{aligned}$$

in such a way that

$$(6.11) \quad \eta_t^{(n)} \equiv \frac{1}{n} \sum_{i=1}^n \delta_{Y_i^{(n)}(t)} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i(t)}$$

for all $t \geq 0$. ($Y^{(n)}$ can be thought of as a variant of a Moran model in which mutation takes place continuously rather than only at birth/death events. Note that by the obvious symmetry of $Y^{(n)}$, this coupling also gives the exchangeability of $X(t)$.) It is easy to check that the sequence of empirical measure processes $\{\eta_t^{(n)}, t \geq 0\}$ converges to the Fleming–Viot process. \square

An explicit construction of the particle system can be given in terms of a collection of unit Poisson processes $\{N_{ij}, 1 \leq i < j < \infty\}$ where N_{ij} determines the times at which the j th level looks down to the i th level, and a collection of independent random mappings $U_{jk} : E \times [0, \infty) \mapsto E$ that determine the values at the j th level between the k th and $(k+1)$ th look-downs from the j th level. That is, for $x \in E$ and $j, k \geq 0$, $U_{jk}(x, \cdot)$ is a version of the mutation process starting from x , $X_1(t) = U_{10}(X_1(0), t)$ for all $t \geq 0$, $X_j(t) = U_{j0}(X_j(0), t)$ until the first look-down from level j , and if the k th look-down from the j th level occurs at time τ and the look-down is to level i , then up until the time of the $(k+1)$ th look-down from level j , $X_j(t) = U_{jk}(X_i(\tau), t - \tau)$.

If the mutation process is a diffusion, then the particle system can be obtained as the solution of an infinite system of stochastic differential equations. Let $\{N_{ij}\}$ be as above, and let $\{W_j\}$ be a sequence of independent Brownian motions which are independent of the sequence $\{N_{ij}\}$. Then we can take

$$(6.12) \quad \begin{aligned} X_j(t) = X_j(0) &+ \int_0^t \sigma(X_j(s)) dW_j(s) + \int_0^t b(X_j(s)) ds \\ &+ \sum_{1 \leq i < j} \int_0^t (X_i(s-) - X_j(s-)) dN_{ij}(s). \end{aligned}$$

Note that the solution of this system exists and is unique if, for example, σ and b are Lipschitz continuous.

If the mutation process has a stationary distribution, then we can assume that $X_1(t)$ is defined for all $-\infty < t < \infty$, and taking the processes N_{ij} to be unit Poisson point processes on $(-\infty, \infty)$ ($N_{ij}(a, b]$ will denote the number of points in the interval $(a, b]$), an analogue of the above construction can be carried out on the time interval $(-\infty, \infty)$ in which there is a look-down from level j to level i at each point in N_{ij} . Note that the resulting E^∞ -valued process will be stationary. If the stationary process X_1 is ergodic (in the sense that events defined in terms of X_1 that are invariant under a time shift must have probability zero or one), then X (and hence the corresponding Fleming–Viot process) is ergodic.

Assume that we are in the stationary setting described in the preceding paragraph. In this model, we can trace the ancestry of a particle by following the look-down process backward in time. For $s < t$, we define $a_j(s, t)$ to be the level of the ancestor at time s of the j th-level particle at time t . To be precise, for $s < t$, let $N_j(s, t] = \sum_{i:i < j} N_{ij}(s, t]$. Define $\gamma_j(t) = \sup\{u < t : N_j(u, t] > 0\}$ and let $\alpha_j(\gamma_j(t))$ be the index i such that $\gamma_j(t) \in N_{ij}$. Define $a_j(s, t) = j$ for $\gamma_j(t) \leq s < t$ and $a_j(s, t) = \alpha_j(\gamma_j(t))$ for $\gamma_{\alpha_j(\gamma_j(t))}(\gamma_j(t)) \leq s < \gamma_j(t)$, and extend the definition of $a_j(s, t)$ to all $s < t$ in the obvious manner.

Let $\Gamma_n(s, t) = \{a_j(s, t) : j = 1, \dots, n\}$ (allowing $n = \infty$). Then, letting $|\Gamma_n(s, t)|$ denote the cardinality of $\Gamma_n(s, t)$, it follows that for $s < t$, $|\Gamma_\infty(s, t)| < \infty$. In particular, $D_n(u) = |\Gamma_n(t - u, t)|$ is a pure death process with transition intensity $k(k - 1)/2$ from state k . If we define the equivalence relation R_u on $\{1, 2, \dots\}$ by setting $i \sim j$ if $a_i(t - u, t) = a_j(t - u, t)$, then $\{R_u, u \geq 0\}$ is Kingman's (1982b) coalescent.

For definiteness, fix $t = 0$ in the above construction. Define $\tau_1 = \inf\{u : D_\infty(u) = 1\}$. Then this construction determines an embedding of a tree in the particle system on the time interval $[-\tau_1, 0]$ whose branches are formed by the ancestral lines of the particles at time 0. The mutation processes determine the evolution of the types of the particles beginning with the type of the two particles at time $-\tau_1$ with descendants at time zero. (Note that one of the two particles will be X_1 .) This observation demonstrates that the genealogical tree with mutation considered by Donnelly and Tavaré (1987) can be embedded in the particle system.

This embedding of the genealogical tree in the particle system provides a powerful tool for studying Fleming-Viot processes. For example, Theorem 7.2 below can be proved by observing that, for pure-jump mutation operators, the genealogical structure implies that, except at points of discontinuity of a component X_i , $X_i = X_j$ for infinitely many j . This type of argument is also at the heart of the results of Dawson and Hochberg (1982) on the support properties of the Fleming-Viot process with Brownian mutation.

7. Bounded mutation operators. Here we consider Fleming-Viot processes with mutation operators A on $B(E)$ of the form

$$(7.1) \quad (Af)(x) = \frac{1}{2}\theta(x) \int_E (f(\xi) - f(x)) P(x, d\xi),$$

where $\theta \in B(E)$ is nonnegative and $P(x, d\xi)$ is a one-step transition function on $E \times B(E)$. (The factor $\frac{1}{2}$ is of course unnecessary but will be convenient later.)

THEOREM 7.1. *Let \mathcal{L} be the generator (3.20) and (3.21) for the Fleming-Viot process with type space E , mutation operator A as in (7.1), no recombination, and selection intensity function $\sigma \in B_{\text{sym}}(E^2)$.*

- (a) *Then the $C_{\mathcal{P}(E)}[0, \infty)$ martingale problem for \mathcal{L} is well posed.*
- (b) *In the absence of selection, the closure of \mathcal{L} acting on*

$$(7.2) \quad \{\varphi_f \in B(\mathcal{P}(E)) : f \in B(E^n), n \geq 1\}$$

generates a strongly continuous positive conservative contraction semigroup on the closure of (7.2) in the sup norm on $B(\mathcal{P}(E))$.

- (c) *In the absence of selection, and assuming that E is compact, θ is continuous, and $P(x, d\xi)$ has the Feller property, the closure of \mathcal{L} restricted to*

$$(7.3) \quad \{\varphi_f \in C(\mathcal{P}(E)) : f \in C(E^n), n \geq 1\}$$

generates a Feller semigroup on $C(\mathcal{P}(E))$.

Proof. Part (c) follows from Theorem 3.4, and part (b) is proved similarly. This and the Girsanov transformation give existence in part (a), while uniqueness is a consequence of duality. \square

The main effect of assumption (7.1) is that the resulting Fleming-Viot process lives in $\mathcal{P}_a(E)$, the set of purely atomic Borel probability measures on E . This result

is due to Ethier and Kurtz (1986), (1987). It is also a consequence of the work of Shiga (1990).

THEOREM 7.2. *If $\{\mu_t, t \geq 0\}$ is a Fleming–Viot process as in Theorem 7.1, then*

$$(7.4) \quad \mathbf{P}\{\mu_t \in \mathcal{P}_a(E) \text{ for all } t > 0\} = 1.$$

If Π is a stationary distribution for such a Fleming–Viot process, then $\Pi(\mathcal{P}_a(E)) = 1$.

In fact this result also holds with recombination, but the simple proof of Ethier and Kurtz (1987) does not apply in that case. Donnelly and Kurtz (1993) have extended the theorem to general pure-jump mutation operators.

The next observation is frequently useful in applications. It is a consequence of Doob’s martingale inequality.

PROPOSITION 7.3. *Under the assumptions of Theorem 7.1, if $\{\mu_t, t \geq 0\}$ is a solution of the $C_{\mathcal{P}(E)}[0, \infty)$ martingale problem for \mathcal{L} , then for each φ in (7.2),*

$$(7.5) \quad \varphi(\mu_t) - \int_0^t (\mathcal{L}\varphi)(\mu_s) ds$$

is a martingale with almost all sample paths continuous.

This shows that the sample paths of such a Fleming–Viot process are continuous in the weak atomic topology on $\mathcal{P}(E)$.

COROLLARY 7.4. *Under the assumptions of Theorem 7.1, if $\{\mu_t, t \geq 0\}$ is a solution of the $C_{\mathcal{P}(E)}[0, \infty)$ martingale problem for \mathcal{L} , then almost all sample paths of $\{\mu_t, t \geq 0\}$ belong to $C_{(\mathcal{P}(E), \rho_a)}[0, \infty)$.*

Shiga (1990) strengthened this conclusion (except at time $t = 0$) considerably.

THEOREM 7.5. *Let E be locally compact, and assume the conditions of Theorem 7.1. If $\{\mu_t, t \geq 0\}$ is a solution of the $C_{\mathcal{P}(E)}[0, \infty)$ martingale problem for \mathcal{L} , then almost all sample paths of $\{\mu_t, t > 0\}$ are continuous in the total variation norm.*

Note that, if $\mu \in \mathcal{P}(E)$ is nonatomic and $\nu \in \mathcal{P}_a(E)$, then $\|\mu - \nu\|_{\text{var}} = 1$. This implies that continuity in the total variation norm cannot hold at time $t = 0$ unless $\mu_0 \in \mathcal{P}_a(E)$.

We conclude this section by discussing the stationary distribution of a Fleming–Viot process with a bounded mutation operator (and no recombination or selection). For simplicity, we assume that E is compact, θ in (7.1) is a positive constant, and $P(x, d\xi)$ has the Feller property. The following result is due to Ethier and Kurtz (1992) and has been used in the population genetics literature in connection with the model discussed in §9.5 below (Griffiths and Watterson (1990)).

THEOREM 7.6. *Let E be compact, let $\theta > 0$, and let $P(x, d\xi)$ be a one-step Feller transition function on $E \times \mathcal{B}(E)$ that is weakly ergodic in the sense that*

$$(7.6) \quad \lim_{t \rightarrow \infty} (e^{\theta(P-I)t/2} f)(x) = \langle f, \nu_0 \rangle, \quad f \in C(E), x \in E,$$

where $(Pf)(x) = \int_E f(\xi) P(x, d\xi)$ and $\nu_0 \in \mathcal{P}(E)$. Let $\Pi \in \mathcal{P}(\mathcal{P}(E))$ denote the unique stationary distribution of the Fleming–Viot process with type space E and mutation operator A defined by

$$(7.7) \quad (Af)(x) = \frac{1}{2}\theta \int_E (f(\xi) - f(x)) P(x, d\xi).$$

Define the Markov chain $\{X(\tau), \tau \in \mathbf{Z}_+\}$ in $E^2 \cup E^3 \cup \dots$ as follows. Let $X(0) = (\xi_0, \xi_0) \in E^2$, where ξ_0 is an E -valued random variable with distribution ν_0 . From

state $(x_1, \dots, x_n) \in E^n$, where $n \geq 2$, one of two types of transitions occurs. With probability $\theta/(n(n-1+\theta))$ a transition to state $(x_1, \dots, x_{i-1}, \xi_i, x_{i+1}, \dots, x_n) \in E^n$ occurs ($1 \leq i \leq n$), where ξ_i is distributed according to $P(x_i, d\xi)$. With probability $(n-1)/((n+1)n(n-1+\theta))$ a transition to state $(x_1, \dots, x_{j-1}, x_i, x_j, \dots, x_n) \in E^{n+1}$ occurs ($1 \leq i \leq n$, $1 \leq j \leq n+1$). Then, for each $n \geq 2$,

$$(7.8) \quad X(\tau_{n+1} - 1) \text{ has distribution } \int_{\mathcal{P}(E)} \mu^n(\cdot) \Pi(d\mu),$$

where τ_n denotes the hitting time of E^n . Moreover, the empirical measure determined by the n coordinates of $X(\tau_{n+1} - 1)$ converges almost surely as $n \rightarrow \infty$ to a $\mathcal{P}(E)$ -valued random variable with distribution Π .

The two types of transitions of X correspond to mutations and duplications, which is reminiscent of Hoppe's (1987) urn process. Equation (7.8) makes it very easy to simulate the n th moment measure of the stationary distribution Π , which can be interpreted as the distribution of a random sample of size n from μ , where the random measure μ is distributed according to Π . Except in the special case in which $P(x, d\xi)$ does not depend on x (treated in the next section), the explicit form of Π seems to be unknown. A continuous-time version of this Markov chain can be embedded in the particle system of §6, giving an extension of (7.8) to models with general mutation operators. See Donnelly and Kurtz (1993).

8. Reversibility. In this section we consider the special case of §7 in which the mutation operator A has the form

$$(8.1) \quad (Af)(x) = \frac{1}{2}\theta \int_E (f(\xi) - f(x)) \nu_0(d\xi),$$

where $\theta > 0$ and $\nu_0 \in \mathcal{P}(E)$. We assume for convenience that E is a compact metric space.

If we further assume that E is finite and the support of ν_0 is all of E , and if we identify $\mathcal{P}(E)$ with Δ_E (see (1.1)), then the resulting Fleming-Viot process has generator L as in (1.2), where

$$(8.2) \quad q_{ij} = \frac{1}{2}\theta_j \equiv \frac{1}{2}\theta\nu_0(\{j\}) > 0, \quad i, j \in E, \quad i \neq j.$$

In this case Wright (1949) discovered that there is a unique stationary distribution $\pi \in \mathcal{P}(\Delta_E)$, namely, the Dirichlet distribution with parameter $(\theta_i)_{i \in E}$, which has $(|E| - 1)$ -dimensional Lebesgue density proportional to $\prod_{i \in E} p_i^{\theta_i - 1}$. See Ethier and Kurtz (1981) for a proof.

To describe the stationary distribution (unique by Theorem 5.1) in general, we need to introduce Kingman's (1975) Poisson-Dirichlet distribution. Consider an inhomogeneous Poisson point process on $(0, \infty)$ with intensity function $\theta u^{-1}e^{-u}$ ($u > 0$). With probability 1, the points can be arranged in decreasing order $Z_1 > Z_2 > \dots$ and have a finite sum $Z = Z_1 + Z_2 + \dots$. We define the Poisson-Dirichlet distribution with parameter θ to be the distribution of $(Z_1/Z, Z_2/Z, \dots)$. In particular, it is concentrated on

$$(8.3) \quad \nabla_\infty = \left\{ p = (p_1, p_2, \dots) : p_1 \geq p_2 \geq \dots \geq 0, \sum_{i=1}^{\infty} p_i = 1 \right\},$$

which is topologized as a subset of the product space $[0, 1]^\infty$.

There is an alternative characterization that helps to explain the term “Poisson–Dirichlet” (Kingman (1975)). For each $n \geq 2$, let $(\rho_1^n, \dots, \rho_n^n)$ be $\Delta_{\{1, \dots, n\}}$ -valued with the symmetric Dirichlet distribution with parameter $p_n > 0$. Let $\rho_{(1)}^n \geq \dots \geq \rho_{(n)}^n$ denote the descending order statistics. If $n p_n \rightarrow \theta > 0$, then the distribution of $(\rho_{(1)}^n, \dots, \rho_{(n)}^n, 0, 0, \dots)$ converges weakly on ∇_∞ to the Poisson–Dirichlet distribution with parameter θ .

The following description of the stationary distribution is due to Ethier and Kurtz (1986, 1993a).

THEOREM 8.1. *The Fleming–Viot process with type space E and mutation operator A defined by (8.1) (where $\theta > 0$ and $\nu_0 \in \mathcal{P}(E)$), has a unique stationary distribution $\Pi_{\theta, \nu_0} \in \mathcal{P}(\mathcal{P}(E))$, which is given by*

$$(8.4) \quad \Pi_{\theta, \nu_0}(\cdot) = \mathbf{P} \left\{ \sum_{i=1}^{\infty} \rho_i \delta_{\xi_i} \in \cdot \right\},$$

where (ρ_1, ρ_2, \dots) has the Poisson–Dirichlet distribution with parameter θ , and ξ_1, ξ_2, \dots are i.i.d. ν_0 , independent of (ρ_1, ρ_2, \dots) .

Note that this conclusion is consistent with Theorem 7.2.

The following result of Shiga (1990) and Ethier (1990a) gives reversibility.

THEOREM 8.2. *The Fleming–Viot process of Theorem 8.1 is reversible with respect to Π_{θ, ν_0} . In other words, if $\{\mu_t, -\infty < t < \infty\}$ is a stationary Fleming–Viot process with type space and mutation operator as in Theorem 8.1, then*

$$(8.5) \quad \{\mu_t, -\infty < t < \infty\} \stackrel{\mathcal{D}}{=} \{\mu_{-t}, -\infty < t < \infty\}.$$

One might ask whether reversibility of Fleming–Viot processes holds more generally than for mutation operators A of the form (8.1) (assuming no recombination or selection). When E is finite, Shiga (personal communication) has shown that the answer to this question is negative. It seems reasonable to conjecture that the answer is negative in general.

We now state a recent result of Ethier and Griffiths (1993) giving the transition function for the Fleming–Viot process of Theorem 8.1. It generalizes the finite-dimensional results of Shimakura (1977, 1981) and Griffiths (1979). For each $n \geq 1$ define $\eta_n : E^n \mapsto \mathcal{P}(E)$ by

$$(8.6) \quad \eta_n(x_1, \dots, x_n) = n^{-1}(\delta_{x_1} + \dots + \delta_{x_n}).$$

Let $\{D_t, t \geq 0\}$ be the pure death process in $\mathbf{Z}_+ \cup \{\infty\}$ starting at ∞ with death rates

$$(8.7) \quad \lambda_n = n(n-1+\theta)/2, \quad n \geq 0,$$

(∞ is an entrance boundary) and define

$$(8.8) \quad d_n(t) = \mathbf{P}\{D_t = n\}, \quad n \geq 0, t > 0.$$

It can be shown (see Tavaré (1984)) that

$$(8.9) \quad d_n(t) = \begin{cases} 1 - \sum_{m=1}^{\infty} (2m-1+\theta)(m!)^{-1}(-1)^{m-1}\theta_{(m-1)}e^{-\lambda_m t} & \text{if } n = 0, \\ \sum_{m=n}^{\infty} (2m-1+\theta)(m!)^{-1}(-1)^{m-n}\binom{m}{n}(n+\theta)_{(m-1)}e^{-\lambda_m t} & \text{if } n \geq 1. \end{cases}$$

Here we use the notation $a_{(0)} = 1$ and $a_{(n)} = a(a+1)\cdots(a+n-1)$ for $n \geq 1$.

THEOREM 8.3. *The Fleming–Viot process of Theorem 8.1 has transition function $P(t, \mu, d\nu)$ given for each $t > 0$ and $\mu \in \mathcal{P}(E)$ by*

$$(8.10) \quad \begin{aligned} P(t, \mu, \cdot) &= d_0(t) \Pi_{\theta, \nu_0}(\cdot) \\ &\quad + \sum_{n=1}^{\infty} d_n(t) \int_{E^n} \mu^n(dx_1 \times \cdots \times dx_n) \Pi_{n+\theta, (n+\theta)^{-1}\{n\eta_n(x_1, \dots, x_n) + \theta\nu_0\}}(\cdot). \end{aligned}$$

In particular, the theorem shows that, for each $t > 0$ and $\mu \in \mathcal{P}(E)$, $P(t, \mu, \cdot)$ is a mixture of probability distributions of the form (8.4). This is of course consistent with Theorem 7.2.

Shiga (1990) proved strong ergodicity under these assumptions. The theorem allows one to say more.

COROLLARY 8.4. *Under the assumptions of Theorem 8.3 and for each $\mu \in \mathcal{P}(E)$,*

$$(8.11) \quad \|P(t, \mu, \cdot) - \Pi_{\theta, \nu_0}(\cdot)\|_{\text{var}} \leq 1 - d_0(t), \quad t > 0.$$

Moreover, equality holds in (8.11) if μ and ν_0 are mutually singular.

Tavaré (1984) has shown that $e^{-\lambda_1 t} \leq 1 - d_0(t) \leq (1 + \theta)e^{-\lambda_1 t}$ for all $t > 0$.

We conclude this section by considering the effect of selection on Theorems 8.1 and 8.2. These results are from Ethier and Kurtz (1993a), (1993b).

THEOREM 8.5. *Let $\sigma \in B_{\text{sym}}(E^2)$. Then $\Pi_{\theta, \nu_0}^\sigma \in \mathcal{P}(\mathcal{P}(E))$, defined for the appropriate constant C by*

$$(8.12) \quad \Pi_{\theta, \nu_0}^\sigma(d\mu) = C e^{\langle \sigma, \mu^2 \rangle} \Pi_{\theta, \nu_0}(d\mu),$$

is the unique stationary distribution for the Fleming–Viot process with type space E , mutation operator A defined by (8.1) (where $\theta > 0$ and $\nu_0 \in \mathcal{P}(E)$), and selection intensity function σ .

THEOREM 8.6. *The Fleming–Viot process of Theorem 8.5 is reversible with respect to $\Pi_{\theta, \nu_0}^\sigma$.*

The proof of reversibility is based on the reversibility-preserving transformation of Fukushima and Stroock (1986).

9. Examples. In this section we examine several Fleming–Viot processes of interest in population genetics. To provide a better understanding of the models, a number of proofs are included.

Handa (1990) and Vaillancourt (1990a), (1990b) have considered interacting systems of Fleming–Viot processes to study the effects of geographical structure. Although these processes do not strictly fit into the present framework and are therefore not discussed below, they are very much in the same spirit.

9.1. Continuous-state stepwise-mutation model. Here $E = \mathbf{R}^d$ and

$$(9.1.1) \quad A = \frac{1}{2} \theta \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}, \quad \mathcal{D}(A) = C_c^2(\mathbf{R}^d),$$

for some $\theta > 0$. The resulting Fleming–Viot process is the one originally studied by Fleming and Viot (1979) and extensively analyzed by Dawson and Hochberg (1982).

The discrete-state stepwise-mutation model (with $E = \mathbf{Z}^d$) is due to Ohta and Kimura (1973) and was studied by Shiga (1982).

We consider first the ergodic behavior of the process. The mutation process is d -dimensional Brownian motion (multiplied by a constant), so the ergodic theorems of §5 do not apply. Nevertheless, if we translate the random probability measure at time t by minus its empirical mean, convergence in distribution holds as $t \rightarrow \infty$. This extends a result of Dawson and Hochberg (1982) and is analogous to a result of Shiga (1982) for the discrete-state stepwise-mutation model. It has its origins in work of Moran (1975), (1976) and Kingman (1976).

The infinite particle system X discussed in §6 provides an alternative approach to this result. Centering each component of X by the first component X_1 instead of by the empirical mean, the resulting process $\tilde{X} = (0, X_2 - X_1, X_3 - X_1, \dots)$ is easily seen to be asymptotically stationary. See Donnelly and Kurtz (1993) for details.

To prove the theorem, we first need a lemma.

LEMMA 9.1.1. *Let $\{\nu_n\}$ be a sequence of $\mathcal{P}(\mathbf{R}^d)$ -valued random variables such that, for each $n \geq 1$ and $m \geq 1$, $\mathbf{E}[\int_{\mathbf{R}^d} |x|^m \nu_n(dx)] < \infty$. If*

$$(9.1.2) \quad \lim_{n \rightarrow \infty} \mathbf{E} \left[\int_{\mathbf{R}^d} x^{\alpha^1} \nu_n(dx) \cdots \int_{\mathbf{R}^d} x^{\alpha^k} \nu_n(dx) \right] \equiv r_k(\alpha^1, \dots, \alpha^k)$$

exists for all $k \geq 1$ and $\alpha^1, \dots, \alpha^k \in (\mathbf{Z}_+)^d$, where $x^\alpha = x_1^{\alpha^1} \cdots x_d^{\alpha^d}$, and if

$$(9.1.3) \quad \sup_{m \geq 1} \max_{1 \leq i \leq d} m^{-1} r_1(2m\varepsilon^i)^{1/(2m)} < \infty,$$

where $\varepsilon^i = (\delta_{i1}, \dots, \delta_{id})$, then there exists a $\mathcal{P}(\mathbf{R}^d)$ -valued random variable ν such that ν_n converges in distribution to ν as $n \rightarrow \infty$.

Proof. By (9.1.2) with $k = 1$, $\{\mathbf{E}[\nu_n(\cdot)]\}$ is tight, and hence tightness (and therefore relative compactness) of the sequence $\mathbf{P}\{\nu_n \in \cdot\}$ follows. Any limit point Π of this sequence has the property that its k th moment measure $\int_{(\mathbf{R}^d)^k} \mu^k(\cdot) \Pi(d\mu)$ has moments $r_k(\alpha^1, \dots, \alpha^k)$. By (9.1.3), these moment measures are uniquely determined, and consequently Π is uniquely determined. This implies that $\mathbf{P}\{\nu_n \in \cdot\} \Rightarrow \Pi$. \square

THEOREM 9.1.2. *Let $\{\mu_t, t \geq 0\}$ be a Fleming–Viot process with type space \mathbf{R}^d and mutation operator (9.1.1), and suppose that $\mathbf{E}[\int_{\mathbf{R}^d} |x|^m \mu_0(dx)] < \infty$ for each $m \geq 1$. Then*

$$(9.1.4) \quad \mathbf{E} \left[\sup_{0 \leq t \leq T} \int_{\mathbf{R}^d} |x|^m \mu_t(dx) \right] < \infty$$

for all $T > 0$ and $m \geq 1$, so, with probability 1, the empirical mean process

$$(9.1.5) \quad x(t) = \int_{\mathbf{R}^d} x \mu_t(dx), \quad t \geq 0,$$

exists and has continuous sample paths. Let $\{\mu_t^, t \geq 0\}$ denote the centered process, defined by*

$$(9.1.6) \quad \mu_t^*(\cdot) = \mu_t(\cdot + x(t)).$$

Then there exists $\Pi \in \mathcal{P}(\mathcal{P}(\mathbf{R}^d))$ not depending on μ_0 such that

$$(9.1.7) \quad \lim_{t \rightarrow \infty} \mathbf{E}[\varphi(\mu_t^*)] = \langle \varphi, \Pi \rangle, \quad \varphi \in \overline{C}(\mathcal{P}(\mathbf{R}^d)).$$

Proof. By Theorem 3.1 (with a slight change in notation),

$$(9.1.8) \quad \mathbf{E}[\langle f, \mu_t^l \rangle] = \mathbf{E}[\langle Y(t), \mu_0^{N(t)} \rangle]$$

for all $l \geq 1$, $f \in \overline{C}((\mathbf{R}^d)^l)$, and $t \geq 0$, where

$$(9.1.9) \quad Y(t) = T_k \left(t - \sum_{i=k+1}^l \Lambda_i \right) \Gamma_{k+1} \cdots \Gamma_{l-1} T_{l-1}(\Lambda_{l-1}) \Gamma_l T_l(\Lambda_l) f$$

and $N(t) = k \quad \text{if} \quad \sum_{i=k+1}^l \Lambda_i \leq t < \sum_{i=k}^l \Lambda_i, \quad 1 \leq k \leq l.$

Here $\Lambda_l, \dots, \Lambda_2$ are independent exponential random variables with $\mathbf{E}[\Lambda_k] = 1/\binom{k}{2}$ and $\Lambda_1 \equiv \infty$,

$$(9.1.10) \quad T_k(t)f(x) = (2\pi\theta t)^{-kd/2} \int_{(\mathbf{R}^d)^k} f(y) e^{-|y-x|^2/(2\theta t)} dy,$$

and $\Gamma_l, \dots, \Gamma_2$ are independent random operators with $\Gamma_k = \Phi_{ij}^{(k)}$ with probability $1/\binom{k}{2}$, $1 \leq i < j \leq k$. Furthermore, $\Lambda_l, \dots, \Lambda_2, \Gamma_l, \dots, \Gamma_2$, and μ_0 are independent.

In fact, we claim that (9.1.8) holds (with both sides finite) for all $l \geq 1$, $f \in C((\mathbf{R}^d)^l)$ with polynomial growth, and $t \geq 0$. Indeed, if we define $f_{m,k}$ on $(\mathbf{R}^d)^k$ for $m, k \geq 1$ by

$$(9.1.11) \quad f_{m,k}(x^1, \dots, x^k) = \left(1 + \sum_{i=1}^k |x^i|^2 \right)^m, \quad x^1, \dots, x^k \in \mathbf{R}^d,$$

then straightforward estimates give

$$(9.1.12) \quad T_k(t)f_{m,k} \leq C_{m,k}(t)f_{m,k} \quad \text{and} \quad \Gamma_k f_{m,k} \leq 2^m f_{m,k-1},$$

where $C_{m,k}(t) = 2^{2m-1}(1 + (kd\theta/2)^m(2m)!(m!)^{-1}t^m)$. Therefore,

$$(9.1.13) \quad |Y(t)| \leq 2^{(l-k)m} C_{m,l}(\Lambda_l) \cdots C_{m,k+1}(\Lambda_{k+1}) C_{m,k} \left(t - \sum_{i=k+1}^l \Lambda_i \right) f_{m,k}$$

if $|Y(0)| \leq f_{m,l}, \quad \sum_{i=k+1}^l \Lambda_i \leq t < \sum_{i=k}^l \Lambda_i, \quad 1 \leq k \leq l.$

Since exponential random variables have all moments finite, this implies the claimed assertion. Doob's martingale inequality applied to the continuous martingale

$$(9.1.14) \quad M_t^\varepsilon \equiv \langle f_{m,1}/(1 + \varepsilon f_{m,1}), \mu_t \rangle - \int_0^t \langle A[f_{m,1}/(1 + \varepsilon f_{m,1})], \mu_s \rangle ds,$$

leads to (9.1.4).

Now, given $\alpha \in (\mathbf{Z}_+)^d$, define $g_\alpha \in C((\mathbf{R}^d)^{|\alpha|+1})$ by

$$(9.1.15) \quad g_\alpha(y^0, y^1, \dots, y^{|\alpha|}) = \prod_{i=1}^d \prod_{j=\alpha_1+\dots+\alpha_{i-1}+1}^{\alpha_1+\dots+\alpha_i} (y_i^0 - y_i^j),$$

and note that g_α is translation invariant in the sense that

$$(9.1.16) \quad g_\alpha(y^0 + z, y^1 + z, \dots, y^{|\alpha|} + z) = g_\alpha(y^0, y^1, \dots, y^{|\alpha|})$$

for all $y^0, \dots, y^{|\alpha|}, z \in \mathbf{R}^d$. Since translation invariance is preserved by the semigroup (9.1.10) and by the operators $\Phi_{ij}^{(k)}$, the dual process $\{Y_\alpha(t), t \geq 0\}$ with $Y_\alpha(0) = g_\alpha$ is translation invariant valued. Let $\tau = \Lambda_{|\alpha|+1} + \dots + \Lambda_2$. Then $Y_\alpha(\tau)$ is a (random) translation invariant function of one variable ($x \in \mathbf{R}^d$), hence constant in x . Moreover,

$$(9.1.17) \quad \mathbf{E} \left[\int_{\mathbf{R}^d} x^\alpha \mu_t^*(dx) \right] = \mathbf{E}[\langle g_\alpha, \mu_t^{|\alpha|+1} \rangle] = \mathbf{E}[\langle Y_\alpha(t), \mu_0^{N(t)} \rangle] \\ \rightarrow \mathbf{E}[\langle Y_\alpha(\tau), \mu_0 \rangle] = \mathbf{E}[Y_\alpha(\tau)(0)]$$

as $t \rightarrow \infty$, where the limit is justified using (9.1.13). This gives (9.1.2) (with $\nu_n = \mu_{t_n}^*$) for $k = 1$, and the general case is similar. Since

$$(9.1.18) \quad \mathbf{E} \left[\int_{\mathbf{R}^d} x_i^{2m} \mu_t^*(dx) \right] = \mathbf{E} \left[\int_{\mathbf{R}^d} (x_i - x_i(t))^{2m} \mu_t(dx) \right] \\ \leq \mathbf{E} \left[\int_{\mathbf{R}^d} \int_{\mathbf{R}^d} (x_i - y_i)^{2m} \mu_t(dx) \mu_t(dy) \right] \\ = \mathbf{E}[\langle Y(t), \mu_0^{N(t)} \rangle] \\ \rightarrow \mathbf{E}[Y(\tau)(0)] \\ = \int_0^\infty e^{-t} \mathbf{E}[(B_1(t) - B_2(t))^{2m}] dt \\ = \theta^m (2m)!$$

for $i = 1, \dots, d$, where τ is exponential with parameter 1 and B_1 and B_2 are independent one-dimensional Brownian motions with variance parameter θ , (9.1.3) holds as well, and so the result follows from Lemma 9.1.1. \square

The most striking results about this process concern its sample path behavior. Here is a small part of what is known. The first conclusion is due to Konno and Shiga (1988) and Reimers (1989) and the second is due to Dawson and Hochberg (1982) and Reimers (1993); see also Shiga (1990).

THEOREM 9.1.3. *Let $\{\mu_t, t \geq 0\}$ be a Fleming–Viot process with type space \mathbf{R}^d and mutation operator (9.1.1).*

(a) *If $d = 1$, then, with probability 1, μ_t is absolutely continuous with respect to Lebesgue measure on \mathbf{R} for all $t > 0$ and its density is jointly continuous in t and x .*

(b) *If $d \geq 2$, then, with probability 1, μ_t is singular with respect to Lebesgue measure on \mathbf{R}^d for all $t > 0$.*

9.2. Infinitely-many-neutral-alleles model. Here $E = S$, which is assumed to be a compact metric space, and A is given by

$$(9.2.1) \quad (Af)(x) = \frac{1}{2}\theta \int_S (f(\xi) - f(x)) P(x, d\xi),$$

where $\theta > 0$ and $P(x, d\xi)$ is a one-step Feller transition function on $S \times \mathcal{B}(S)$ satisfying

$$(9.2.2) \quad P(x, \{\xi\}) = 0, \quad x, \xi \in S.$$

In other words, mutations occur at rate $\frac{1}{2}\theta$, and $P(x, \cdot)$ is the distribution of the allelic type of a mutant offspring of a type x parent. The condition (9.2.2) models the basic assumption that every mutant is of a new type.

The model is due to Kimura and Crow (1964); see Crow (1989) for a historical survey. Kingman's (1975) introduction of the Poisson-Dirichlet distribution led Waterson (1976) to formulate the model as the limit of a sequence of finite-dimensional diffusions. Ethier and Kurtz (1981) characterized the limiting process in

$$(9.2.3) \quad \bar{\nabla}_\infty = \left\{ p = (p_1, p_2, \dots) : p_1 \geq p_2 \geq \dots \geq 0, \sum_{i=1}^{\infty} p_i \leq 1 \right\}$$

and subsequently (Ethier and Kurtz (1986), (1987)) reformulated the model as a Fleming-Viot process.

The process in $\bar{\nabla}_\infty$ is characterized in terms of the generator

$$(9.2.4) \quad G = \frac{1}{2} \sum_{i,j=1}^{\infty} p_i (\delta_{ij} - p_j) \frac{\partial^2}{\partial p_i \partial p_j} - \frac{1}{2} \theta \sum_{i=1}^{\infty} p_i \frac{\partial}{\partial p_i}$$

acting on $\mathcal{D}(G) =$ subalgebra of $C(\bar{\nabla}_\infty)$ generated by $1, \varphi_2, \varphi_3, \dots$, where $\varphi_m \in C(\bar{\nabla}_\infty)$ is defined for $m \geq 2$ by $\varphi_m(p) = \sum_{i=1}^{\infty} p_i^m$. (Sums in (9.2.4) are evaluated on ∇_∞ (see (8.3)) and extended to $\bar{\nabla}_\infty$ by continuity.) It is known (Ethier and Kurtz (1981)) that this process has a unique stationary distribution, namely, the Poisson-Dirichlet distribution with parameter θ . (In fact, it has a transition density; see Griffiths (1979) and Ethier (1992b).)

Define $\gamma : \mathcal{P}(S) \mapsto \bar{\nabla}_\infty$ by letting $\gamma(\mu) = (p_1, p_2, \dots)$ if p_i is the size (or mass) of the i th largest atom of μ (or 0 if μ has fewer than i atoms) for each $i \geq 1$, and note that $\gamma(\mathcal{P}_a(S)) = \nabla_\infty$.

THEOREM 9.2.1. *Let $\{\mu_t, t \geq 0\}$ be a Fleming-Viot process with type space S and mutation operator A given by (9.2.1), where $\theta > 0$ and $P(x, d\xi)$ is a one-step Feller transition function on $S \times \mathcal{B}(S)$ satisfying (9.2.2). Then $\{\gamma(\mu_t), t \geq 0\}$ solves the $C_{\bar{\nabla}_\infty}[0, \infty)$ martingale problem for G .*

Proof. Let $k \geq 1$, $m_1, \dots, m_k \geq 2$, and $m = m_1 + \dots + m_k$. Define $f \in B(S^m)$ to be the indicator function of the set of $(x_1, \dots, x_m) \in S^m$ for which $x_1 = \dots = x_{m_1}$, $x_{m_1+1} = \dots = x_{m_1+m_2}$, \dots , $x_{m_1+\dots+m_{k-1}+1} = \dots = x_m$. Then

$$(9.2.5) \quad F(\mu) \equiv \langle f, \mu^m \rangle = \varphi_{m_1}(\gamma(\mu)) \cdots \varphi_{m_k}(\gamma(\mu))$$

and

$$(9.2.6) \quad \begin{aligned} (\mathcal{L}F)(\mu) &= \sum_{i=1}^k \binom{m_i}{2} \varphi_{m_i-1}(\gamma(\mu)) \prod_{l:l \neq i} \varphi_{m_l}(\gamma(\mu)) \\ &\quad + \sum_{1 \leq i < j \leq k} m_i m_j \varphi_{m_i+m_j-1}(\gamma(\mu)) \prod_{l:l \neq i,j} \varphi_{m_l}(\gamma(\mu)) \\ &\quad - \frac{1}{2} m(m-1+\theta) \varphi_{m_1}(\gamma(\mu)) \cdots \varphi_{m_k}(\gamma(\mu)) \\ &= G(\varphi_{m_1} \cdots \varphi_{m_k})(\gamma(\mu)) \end{aligned}$$

for all $\mu \in \mathcal{P}(S)$, where $\varphi_1 \equiv 1$, and the result follows using Proposition 7.3. \square

Both $\{\mu_t, t \geq 0\}$ and $\{\gamma(\mu_t), t \geq 0\}$ are referred to as the *infinitely-many-neutral-alleles diffusion model*; for obvious reasons, the former is sometimes called the *labeled model*, whereas the latter is known as the *unlabeled model*.

A related diffusion process assumes values in

$$(9.2.7) \quad \bar{\Delta}_\infty = \left\{ z = (z_1, z_2, \dots) : z_1 \geq 0, z_2 \geq 0, \dots, \sum_{i=1}^{\infty} z_i \leq 1 \right\}.$$

It was characterized by Ethier (1981) in terms of the generator

$$(9.2.8) \quad \Omega = \frac{1}{2} \sum_{i,j=1}^{\infty} z_i (\delta_{ij} - z_j) \frac{\partial^2}{\partial z_i \partial z_j} - \frac{1}{2} \theta \sum_{i=1}^{\infty} z_i \frac{\partial}{\partial z_i}$$

acting on $\mathcal{D}(\Omega) = \{f \in C^2(\bar{\Delta}_\infty) : f \text{ depends on only finitely many coordinates}\}$. To interpret this process, we define $\varphi : \mathcal{P}(S) \times \mathcal{P}(S) \mapsto \bar{\Delta}_\infty$ by letting $\varphi(\mu, \nu) = (z_1, z_2, \dots)$ if z_i is the mass given by μ to the i th largest atom of ν (or 0 if ν has fewer than i atoms) for each $i \geq 1$. (Technical note: There is an ambiguity in the definition if ν has two or more atoms of the same size. To overcome this, assume that S is totally ordered by $<$ and that $\{(x_1, x_2) \in S^2 : x_1 < x_2\}$ is a Borel set. The atoms of the same size can then be ordered according to $<$, and the mapping φ is Borel measurable.) The following result is a slight extension of a result of Ethier (1990a).

THEOREM 9.2.2. *With $\{\mu_t, t \geq 0\}$ as in Theorem 9.2.1, $\{\varphi(\mu_t, \mu_0), t \geq 0\}$ solves the $C_{\bar{\Delta}_\infty}[0, \infty)$ martingale problem for Ω .*

Note that the i th coordinate of $\varphi(\mu_t, \mu_0)$ is the frequency at time t of the allele that is i th most frequent at time 0. Thus, this process keeps track only of the frequencies of the alleles present at time 0 and ignores new mutants. In particular, $\mathbf{E}[\sum_{i=1}^{\infty} \varphi_i(\mu_t, \mu_0)] \leq e^{-\theta t/2} < 1$ for all $t > 0$.

Now let us define $P : C(S) \mapsto C(S)$ by $(Pf)(x) = \int_S f(\xi) P(x, d\xi)$ and assume that there exists $\nu_0 \in \mathcal{P}(S)$ such that

$$(9.2.9) \quad \lim_{t \rightarrow \infty} (e^{\theta(P-I)t/2} f)(x) = \langle f, \nu_0 \rangle, \quad f \in C(S), x \in S.$$

THEOREM 9.2.3. *The Fleming–Viot process with type space S and mutation operator A given by (9.2.1), where $\theta > 0$ and $P(x, d\xi)$ is a one-step Feller transition function on $S \times \mathcal{B}(S)$ satisfying (9.2.2) and (9.2.9), is weakly ergodic and has a unique stationary distribution $\Pi \in \mathcal{P}(\mathcal{P}(S))$. Moreover, the Π -distribution of γ is precisely the Poisson–Dirichlet distribution with parameter θ . Let $k \geq 1$, $\mathbf{n} = (n_1, \dots, n_k) \in \mathbf{N}^k$, and $n = n_1 + \dots + n_k$, and define*

$$(9.2.10) \quad \Gamma_{\mathbf{n}} = \{(x_1, \dots, x_n) \in S^n : \text{there exist distinct } y_1, \dots, y_k \in S \text{ such that } |\{1 \leq i \leq n : x_i = y_j\}| = n_j \text{ for } j = 1, \dots, k\}.$$

Then the Ewens (1972) sampling formula holds, that is,

$$(9.2.11) \quad \int_{\mathcal{P}(S)} \mu^n(\Gamma_{\mathbf{n}}) \Pi(d\mu) = \frac{n!}{n_1 \cdots n_k a_1! \cdots a_n!} \frac{\theta^{k-1}}{(1+\theta) \cdots (n-1+\theta)},$$

where $a_i = |\{1 \leq j \leq k : n_j = i\}|$ for $i = 1, \dots, n$.

Proof. The first assertion follows from Theorem 5.2. Let $k \geq 1$, $m_1, \dots, m_k \geq 2$, and $m = m_1 + \dots + m_k$, and define $F \in B(\mathcal{P}(S))$ as in (9.2.5). By (9.2.6),

$$(9.2.12) \quad 0 = \int_{\mathcal{P}(S)} (\mathcal{L}F)(\mu) \Pi(d\mu) = \int_{\overline{\nabla}_\infty} G(\varphi_{m_1} \cdots \varphi_{m_k}) d\Pi \gamma^{-1},$$

and this gives the second assertion. The last assertion is a property of the Poisson-Dirichlet distribution (see, e.g., Watterson (1976)). \square

We should emphasize that the results of this subsection rely on (9.2.2); the assumption (8.1) (and in particular reversibility) is not needed.

9.3. Infinitely-many-neutral-alleles diffusion model with ages. Here S , θ , and $P(x, d\xi)$ are as in §9.2, $E = S \times [0, \infty]$, and

$$(9.3.1) \quad (Af)(x, a) = \left(\frac{\partial}{\partial a} f \right)(x, a) + \frac{1}{2} \theta \int_S (f(\xi, 0) - f(x, a)) P(x, d\xi)$$

with $\mathcal{D}(A) = C^{0,1}(S \times [0, \infty])$. We associate an age with each allele represented in the population, and we take the type space E to be the set of all ordered pairs of alleles and ages. Again, mutations occur with intensity $\frac{1}{2}\theta$ and according to $P(x, d\xi)$, and new mutants initially have age 0. Ages increase deterministically with time at rate 1.

Ages of alleles have been studied by several authors, including Watterson and Guess (1977), Kelly (1979), Donnelly (1986), Donnelly and Tavaré (1986), (1987), and Hoppe (1987). The formulation as a Fleming-Viot process is from Ethier (1990a).

The mutation operator (9.3.1) is unbounded, so Theorem 7.2 does not apply. Nevertheless, its conclusion is valid here; in fact, we can say more. Let

$$(9.3.2) \quad \begin{aligned} \mathcal{P}_a^0(S \times [0, \infty]) = \{ \mu \in \mathcal{P}_a(S \times [0, \infty]) : \mu(S \times \{\infty\}) = 0, \\ \mu^2(\{((x_1, a_1), (x_2, a_2)) : (x_1 = x_2, a_1 \neq a_2) \text{ or } \\ (x_1 \neq x_2, a_1 = a_2)\}) = 0 \} \end{aligned}$$

The extra conditions mean that no allele has infinite age, each allele has only one age, and no two alleles have the same age.

LEMMA 9.3.1. *Let $\{\mu_t, t \geq 0\}$ be a Fleming-Viot process with type space $S \times [0, \infty]$ and mutation operator A given by (9.3.1), where $\theta > 0$ and $P(x, d\xi)$ is a Feller transition function satisfying (9.2.2), and suppose that $\mathbf{P}\{\mu_0 \in \mathcal{P}_a^0(S \times [0, \infty])\} = 1$. Then*

$$(9.3.3) \quad \mathbf{P}\{\mu_t \in \mathcal{P}_a^0(S \times [0, \infty]) \text{ for all } t \geq 0\} = 1.$$

We now assume a slightly stronger condition than (9.2.9), namely,

$$(9.3.4) \quad \lim_{n \rightarrow \infty} \int_S f(\xi) P^n(x, d\xi) = \langle f, \nu_0 \rangle, \quad f \in C(S), \quad x \in S,$$

LEMMA 9.3.2. *Assuming (9.3.4), the Fleming-Viot process of Lemma 9.3.1 is weakly ergodic and has a unique stationary distribution $\Pi \in \mathcal{P}(\mathcal{P}(S \times [0, \infty]))$. Also, $\Pi(\mathcal{P}_a^0(S \times [0, \infty])) = 1$.*

The Ewens sampling formula (9.2.12) was generalized to age-ordered samples by Donnelly and Tavaré (1986). Ethier (1990a) rederived the Donnelly-Tavaré sampling formula in the present framework.

THEOREM 9.3.3. *Let $k \geq 1$, $\mathbf{n} = (n_1, \dots, n_k) \in \mathbf{N}^k$, and $n = n_1 + \dots + n_k$, and define*

$$(9.3.5) \quad \Gamma_{\mathbf{n}}^* = \{((x_1, a_1), \dots, (x_n, a_n)) \in (S \times [0, \infty])^n : \\ \text{there exist distinct } y_1, \dots, y_k \in S \text{ and } 0 < \alpha_1 < \dots < \alpha_k < \infty \\ \text{such that } |\{1 \leq i \leq n : (x_i, a_i) = (y_j, \alpha_j)\}| = n_j \text{ for } j = 1, \dots, k\}.$$

Then, under the assumptions of Lemma 9.3.2,

$$(9.3.6) \quad \int_{\mathcal{P}(S \times [0, \infty])} \mu^n(\Gamma_{\mathbf{n}}^*) \Pi(d\mu) \\ = \frac{n!}{n_1(n_1 + n_2) \cdots (n_1 + \dots + n_k)} \frac{\theta^{k-1}}{(1 + \theta) \cdots (n - 1 + \theta)}.$$

Note that $\Gamma_{\mathbf{n}}^*$ is the set of all ordered samples of size n containing k alleles, of which the youngest has n_1 representatives, the second youngest has n_2 representatives, and so on.

We now consider a second way to study the ages of the alleles in the infinitely-many-neutral-alleles diffusion model. It is based on the sample paths of the Fleming–Viot process of §9.2, except that we require reversibility. The following lemma is from Ethier (1990a).

LEMMA 9.3.4. *Let $\{\mu_t, -\infty < t < \infty\}$ be a stationary Fleming–Viot process with type space S and mutation operator*

$$(9.3.7) \quad (Af)(x) = \frac{1}{2}\theta \int_S (f(\xi) - f(x)) \nu_0(d\xi),$$

where $\theta > 0$ and $\nu_0 \in \mathcal{P}(S)$ is nonatomic. Recall the notation in Theorem 9.2.2 and define

$$(9.3.8) \quad X^+(t) = \varphi(\mu_t, \mu_0) \quad \text{and} \quad X^-(t) = \varphi(\mu_{-t}, \mu_0), \quad t \geq 0.$$

Then almost all sample paths of X^+ and X^- belong to $C_{\overline{\Delta}_\infty}[0, \infty)$, and X^+ and X^- induce the same distribution on $C_{\overline{\Delta}_\infty}[0, \infty)$.

Theorem 9.2.2 shows how X^+ (and therefore X^-) evolves; and Theorem 9.2.3 implies that $X^+(0) \equiv X^-(0)$ has the Poisson–Dirichlet distribution with parameter θ . The proof of Lemma 9.3.4 is essentially immediate from the reversibility of $\{\mu_t, -\infty < t < \infty\}$ (Theorem 8.2).

For each $i \geq 1$, define $\tau_i : C_{\overline{\Delta}_\infty}[0, \infty) \mapsto [0, \infty]$ by $\tau_i(x) = \inf\{t \geq 0 : x_i(t) = 0\}$, where $\inf \emptyset = \infty$.

THEOREM 9.3.5. *Let X^- be as in Lemma 9.3.4 and let U be an independent uniform $[0, 1)$ random variable. Define the positive-integer-valued random variables α and β by*

$$(9.3.9) \quad \alpha = i \quad \text{if} \quad \tau_i(X^-) > \tau_j(X^-) \text{ for all } j \in \mathbf{N} - \{i\}$$

and

$$(9.3.10) \quad \beta = i \quad \text{if} \quad \sum_{j=1}^{i-1} X_j^-(0) \leq U < \sum_{j=1}^i X_j^-(0);$$

both are defined on an event of probability 1. Then

$$(9.3.11) \quad X_{\alpha}^{-}(0) \stackrel{\mathcal{D}}{=} X_{\beta}^{-}(0)$$

and the common distribution is the $\text{beta}(1, \theta)$ distribution.

Remark. (9.3.11) says that the frequency of the oldest allele at time 0 in the stationary infinitely-many-neutral-alleles diffusion model is distributed as the frequency of a randomly chosen allele at time 0.

Proof. The assertion (9.3.11) follows from the calculation

$$\begin{aligned} (9.3.12) \quad & \mathbf{P}\{\alpha = i \mid X^{-}(0)\} \\ &= \mathbf{P}\{\tau_i(X^{-}) \geq \sup_{j:j \neq i} \tau_j(X^{-}) \mid X^{-}(0)\} \\ &= \mathbf{P}\left\{\tau_1\left(X_i^{-}, \sum_{j:j \neq i} X_j^{-}, 0, 0, \dots\right) \geq \tau_2\left(X_i^{-}, \sum_{j:j \neq i} X_j^{-}, 0, 0, \dots\right) \mid X^{-}(0)\right\} \\ &= X_i^{-}(0) \\ &= \mathbf{P}\{\beta = i \mid X^{-}(0)\}, \end{aligned}$$

where the third equality uses Theorem 9.2.2, Lemma 9.3.4, and the observation that the function $z_1/(z_1 + z_2)$ is harmonic for Ω . Actually, this function has a singularity at $(0, 0, \dots)$, so to make the argument precise and to verify the first equality in (9.3.12), one must check that if Z is a solution of the $C_{\bar{\Delta}_{\infty}}^{-}[0, \infty)$ martingale problem for Ω starting at $z \in \bar{\Delta}_{\infty}$, and if $z_1 + z_2 > 0$, then $\mathbf{P}\{\tau_1(Z) = \tau_2(Z)\} = 0$. See Ethier (1990a) for details; alternatively, Shiga (personal communication) has pointed out that this follows from his representation (Shiga (1990)) of the neutral diffusion model in population genetics as a normalized time-changed continuous-state branching diffusion.

As for the distribution of (9.3.11), let V have the $\text{beta}(1, \theta)$ distribution. Then, for each $k \geq 1$,

$$\begin{aligned} (9.3.13) \quad \mathbf{E}[X_{\beta}^{-}(0)^k] &= \mathbf{E}\left[\sum_{i=1}^{\infty} X_i^{-}(0)^k \mathbf{P}\{\beta = i \mid X^{-}(0)\}\right] \\ &= \mathbf{E}\left[\sum_{i=1}^{\infty} X_i^{-}(0)^{k+1}\right] = \frac{k!}{(1 + \theta) \cdots (k + \theta)} = \mathbf{E}[V^k], \end{aligned}$$

where the third equality uses the Ewens sampling formula, and the result follows. \square

Theorem 9.3.5 is a special case of a result derived heuristically by Griffiths (unpublished) and subsequently established by Ethier (1990b) (see Sawyer (1977), Donnelly and Tavaré (1986), Hoppe (1987), and Donnelly and Joyce (1991) for closely related results). The general result says that the frequencies of the oldest, second-oldest, third-oldest, \dots alleles in the stationary infinitely-many-neutral-alleles diffusion model are distributed as the coordinates of the so-called size-biased Poisson–Dirichlet distribution with parameter θ , which in turn are distributed as $Z_1, (1 - Z_1)Z_2, (1 - Z_1)(1 - Z_2)Z_3, \dots$, where Z_1, Z_2, Z_3, \dots are independent $\text{beta}(1, \theta)$. This last distribution is now known as the GEM distribution with parameter θ . The equality of the size-biased Poisson–Dirichlet and the GEM is a result of Patil and Taillie (1977) (see Donnelly and Joyce (1989) for a proof). We remark that Theorem 9.3.5 (and the

generalization) can be proved in the generality of Theorem 9.3.3 (see Ethier (1990b), (1992a)), but the present argument using reversibility is more appealing.

9.4. The two-locus model with recombination. Here $E = E_1 \times E_2$, where E_1 and E_2 are for simplicity assumed to be compact metric spaces representing the sets of alleles (or types) at two particular loci. Mutation operates independently at the two loci in accordance with mutation operators A_1 and A_2 , which are assumed to generate Feller semigroups on $C(E_1)$ and $C(E_2)$, respectively. Thus, the mutation operator here is the closure \bar{A} of the linear operator A defined on $\mathcal{D}(A) = \text{span}\{f_1 \times f_2 : f_1 \in \mathcal{D}(A_1), f_2 \in \mathcal{D}(A_2)\}$ by

$$(9.4.1) \quad A(f_1 \times f_2) = A_1 f_1 \times f_2 + f_1 \times A_2 f_2,$$

where $(f_1 \times f_2)(x_1, x_2) = f_1(x_1)f_2(x_2)$.

The distinguishing feature of the model, however, is recombination. Assume that the transition function R_M in (2.1.7) and (2.2.4) has the form, for each $M \geq 1$,

$$(9.4.2) \quad R_M((x_1, x_2), (y_1, y_2), \cdot) = (1 - r_M)(\delta_{(x_1, x_2)} \times \delta_{(y_1, y_2)}) + r_M(\delta_{(x_1, y_2)} \times \delta_{(y_1, x_2)}),$$

where $0 \leq r_M \leq \frac{1}{2}$. The idea is that, with probability r_M , two chromosomes with respective types (x_1, x_2) and (y_1, y_2) split at some point between the two loci and recombine, yielding chromosomes with types (x_1, y_2) and (y_1, x_2) . If $\lim_{M \rightarrow \infty} M r_M = \alpha \in [0, \infty)$, then the recombination operator B in (3.5) has the form

$$(9.4.3) \quad (Bf)((x_1, x_2), (y_1, y_2)) = \alpha(f(x_1, y_2) - f(x_1, x_2)).$$

The resulting Fleming–Viot process has been studied by many authors in the finite-dimensional case and by Ethier and Griffiths (1990a), (1990b) and Hochberg (1991) in general.

Let us define π_1 and π_2 to be the projections of $E = E_1 \times E_2$ onto its two coordinates. If $\{\mu_t, t \geq 0\}$ is a Fleming–Viot process as above (type space E , mutation operator A , recombination operator B), it is immediate that $\{\mu_t \pi_1^{-1}, t \geq 0\}$ and $\{\mu_t \pi_2^{-1}, t \geq 0\}$ are Fleming–Viot processes with type spaces E_1 and E_2 and mutation operators A_1 and A_2 (and no recombination). Thus, the two-locus diffusion model can be regarded as a Markov coupling of two one-locus diffusion models, with the recombination parameter measuring in some sense the degree of independence between the marginal processes. The case $\alpha = 0$ is known as the case of complete linkage, whereas the case $\alpha \rightarrow \infty$ corresponds to no linkage, as the following theorem shows. The result is due to Littler (1972) and Ethier (1979) in the finite-dimensional case, and to Ethier and Griffiths (1990a) in general.

THEOREM 9.4.1. *Let E_1, E_2, E, A_1, A_2, A , and B be as above. Let $\{\mu_t^{(1)}, t \geq 0\}$ and $\{\mu_t^{(2)}, t \geq 0\}$ be independent Fleming–Viot processes with type spaces E_1 and E_2 and mutation operators A_1 and A_2 , and let $\{\mu_{\alpha, t}, t \geq 0\}$ be a Fleming–Viot process with type space E , mutation operator A , and recombination operator B , with the dependence on α made explicit. If $(\mu_{\alpha, 0} \pi_1^{-1}, \mu_{\alpha, 0} \pi_2^{-1}) \Rightarrow (\mu_0^{(1)}, \mu_0^{(2)})$ in $\mathcal{P}(E_1) \times \mathcal{P}(E_2)$ as $\alpha \rightarrow \infty$, then*

$$(9.4.4) \quad \{(\mu_{\alpha, t} \pi_1^{-1}, \mu_{\alpha, t} \pi_2^{-1}), t \geq 0\} \Rightarrow \{(\mu_t^{(1)}, \mu_t^{(2)}), t \geq 0\}$$

in $C_{\mathcal{P}(E_1) \times \mathcal{P}(E_2)}[0, \infty)$.

This result asserts asymptotic independence of the marginal processes as $\alpha \rightarrow \infty$. As might be expected, a similar conclusion holds for stationary distributions. Define $\zeta : \mathcal{P}(E_1) \times \mathcal{P}(E_2) \mapsto \mathcal{P}(E)$ by $\zeta(\mu_1, \mu_2) = \mu_1 \times \mu_2$.

THEOREM 9.4.2. *Assume, in addition to the assumptions of Theorem 9.4.1, that the Feller semigroups $\{T_1(t)\}$ and $\{T_2(t)\}$ generated by A_1 and A_2 are weakly ergodic, that is, there exist $\nu_1 \in \mathcal{P}(E_1)$ and $\nu_2 \in \mathcal{P}(E_2)$ such that*

$$(9.4.5) \quad \lim_{t \rightarrow \infty} T_i(t)f(x) = \langle f, \nu_i \rangle, \quad f \in C(E_i), \quad x \in E_i, \quad i = 1, 2.$$

Let $\Pi^{(1)} \in \mathcal{P}(\mathcal{P}(E_1))$ and $\Pi^{(2)} \in \mathcal{P}(\mathcal{P}(E_2))$ be the unique stationary distributions of the Fleming–Viot processes with type spaces E_1 and E_2 and mutation operators A_1 and A_2 , and let $\Pi_\alpha \in \mathcal{P}(\mathcal{P}(E))$ be a stationary distribution of the Fleming–Viot process with type space E , mutation operator A , and recombination operator B (unique by Ethier and Griffiths (1990a)). Then $\Pi_\alpha \Rightarrow (\Pi^{(1)} \times \Pi^{(2)})\zeta^{-1}$ on $\mathcal{P}(E)$ as $\alpha \rightarrow \infty$.

In particular, the limiting stationary distribution is concentrated on the set of product measures.

We conclude this subsection with a simple but useful observation. (We continue to assume that E_1, E_2, E, A_1, A_2, A , and B are as above.) Write the typical element of E^n as $((x_1^1, x_2^1), \dots, (x_1^n, x_2^n))$. Suppose $a, b, c \geq 0$ and $n = a + b + c \geq 1$. Denote by $C_{a,b,c}(E^n)$ the space of continuous functions f on E^n that do not depend on x_2^1, \dots, x_2^a or on $x_1^{a+1}, \dots, x_1^{a+b}$. Given such an f , let us say that the “monomial” $\varphi_f \in C(\mathcal{P}(E))$ defined by $\varphi_f(\mu) = \langle f, \mu^n \rangle$ has degree $a + b + 2c$. Then the generator \mathcal{L} of the Fleming–Viot process with type space E , mutation operator A , and recombination operator B preserves degree, that is, if the monomial φ_f belongs to $\mathcal{D}(\mathcal{L})$, then $\mathcal{L}\varphi_f$ is a linear combination of monomials, each of degree at most that of φ_f .

Here is an elementary application of this idea. For $i = 1, 2$, let A_i be defined by

$$(9.4.6) \quad (A_i f)(x) = \frac{1}{2} \theta_i \int_{E_i} (f(\xi) - f(x)) P_i(x, d\xi),$$

where $\theta_i > 0$ and $P_i(x, d\xi)$ is a one-step Feller transition function on $E_i \times \mathcal{B}(E_i)$ satisfying $P_i(x, \{\xi\}) = 0$ for all $x, \xi \in E_i$. The Fleming–Viot process with type space E , mutation operator A , and recombination operator B is the two-locus analogue of the infinitely-many-neutral-alleles diffusion model of §9.2. Assume that the Feller semigroups $\{T_1(t)\}$ on $C(E_1)$ and $\{T_2(t)\}$ on $C(E_2)$ generated by A_1 and A_2 satisfy (9.4.5), and let Π_α be as in Theorem 9.4.2 (with $0 \leq \alpha < \infty$). Then (see Strobeck and Morgan (1978), Griffiths (1981), and Ethier and Griffiths (1990b))

$$(9.4.7) \quad \int_{\mathcal{P}(E)} \mu^2(\{((x_1, x_2), (y_1, y_2)) \in E^2 : x_1 = y_1, x_2 = y_2\}) \Pi_\alpha(d\mu) \\ = \frac{(3 + \theta)(6 + \theta) + \alpha(2 + \theta + 2(6 + \theta)\kappa) + 2\alpha^2\kappa}{(1 + \theta)(3 + \theta)(6 + \theta) + \alpha(2 + \theta)(13 + 3\theta) + 2\alpha^2(2 + \theta)},$$

where $\theta = \theta_1 + \theta_2$ and $\kappa = (1 + \theta_1)^{-1} + (1 + \theta_2)^{-1}$. It seems clear that there is no explicit formula in the present context that is analogous to the Ewens sampling formula (9.2.11).

9.5. n -locus model with gene conversion. Fix $n \geq 2$. Here $E = [0, 1]^n$ and A is defined by

$$\begin{aligned}
 (9.5.1) \quad (Af)(x_1, \dots, x_n) \\
 = v \sum_{i=1}^n \left\{ \int_0^1 f(x_1, \dots, x_{i-1}, \xi, x_{i+1}, \dots, x_n) d\xi - f(x_1, \dots, x_n) \right\} \\
 + \lambda \sum_{i \neq j} \{ (\Psi_{ij}^{(n)} f)(x_1, \dots, x_n) - f(x_1, \dots, x_n) \},
 \end{aligned}$$

where $\Psi_{ij}^{(n)} : C([0, 1]^n) \mapsto C([0, 1]^n)$ is defined by letting $\Psi_{ij}^{(n)} f$ be the function obtained from f by replacing x_j by x_i ; here $v > 0$ is the mutation intensity at each locus, and $\lambda > 0$ is the rate at which the allele at locus j is converted to the allele at locus i ($i, j = 1, \dots, n$; $i \neq j$). The model is due to Ohta (1982), (1983), and the present formulation is that of Shimizu (1985), (1990).

LEMMA 9.5.1. *The pure-jump Markov process in $[0, 1]^n$ with generator A given by (9.5.1) is weakly ergodic and therefore has a unique stationary distribution. Consequently, the Fleming–Viot process with type space $[0, 1]^n$ and mutation operator A is weakly ergodic and therefore has a unique stationary distribution $\Pi \in \mathcal{P}(\mathcal{P}([0, 1]^n))$.*

Proof. By Theorem 5.2, we need only verify the first assertion. Let $\{(x_1(t), \dots, x_n(t)), t \geq 0\}$ denote the Markov process with generator A . The functions $f_{m_1, \dots, m_n}(x_1, \dots, x_n) = f_{x_1, \dots, x_n}(m_1, \dots, m_n) = x_1^{m_1} \cdots x_n^{m_n}$ give rise to a dual process $\{(m_1(t), \dots, m_n(t)), t \geq 0\}$ in $(\mathbf{Z}_+)^n$ with transitions

$$(9.5.2) \quad (m_1, \dots, m_n) \mapsto (m_1, \dots, m_{i-1}, 0, m_{i+1}, \dots, m_n)$$

with rate $v/(m_i + 1)$ for $i = 1, \dots, n$, and

$$(9.5.3) \quad (m_1, \dots, m_n) \mapsto (m_1, \dots, m_{i-1}, m_i + m_j, m_{i+1}, \dots, m_{j-1}, 0, m_{j+1}, \dots, m_n)$$

with rate λ for $i, j = 1, \dots, n$, $i \neq j$. This process absorbs at $(0, \dots, 0)$ in finite time with probability 1. Thus, as $t \rightarrow \infty$,

$$\begin{aligned}
 (9.5.4) \quad \mathbf{E}_{(x_1, \dots, x_n)} [x_1(t)^{m_1} \cdots x_n(t)^{m_n}] \\
 = \mathbf{E}_{(m_1, \dots, m_n)} \left[x_1^{m_1(t)} \cdots x_n^{m_n(t)} \exp \left\{ -v \int_0^t \sum_{i=1}^n \frac{m_i(s)}{m_i(s) + 1} ds \right\} \right] \\
 \rightarrow \mathbf{E}_{(m_1, \dots, m_n)} \left[\exp \left\{ -v \int_0^\infty \sum_{i=1}^n \frac{m_i(s)}{m_i(s) + 1} ds \right\} \right]
 \end{aligned}$$

for all $(x_1, \dots, x_n) \in [0, 1]^n$ and $(m_1, \dots, m_n) \in (\mathbf{Z}_+)^n$, and the weak ergodicity follows. \square

The following result is due to Shimizu (1987), (1990).

THEOREM 9.5.2. *Let Π be as in Lemma 9.5.1, and let Π_0 be the unique stationary distribution for the Fleming–Viot process with type space $[0, 1]$ and mutation operator A_0 given by*

$$(9.5.5) \quad (A_0 f)(x) = \frac{1}{2} \theta \int_0^1 (f(\xi) - f(x)) d\xi,$$

where $\theta = v/\lambda$. Then, for each $k \geq 1$ and $\mathbf{n} = (n_1, \dots, n_k) \in \mathbf{N}^k$,

$$(9.5.6) \quad \int_{\mathcal{P}([0, 1]^n)} \mu(\Gamma_{\mathbf{n}}) \Pi(d\mu) = \int_{\mathcal{P}([0, 1])} \nu^n(\Gamma_{\mathbf{n}}) \Pi_0(d\nu),$$

where $\Gamma_{\mathbf{n}}$ is as in (9.2.10) and $n = n_1 + \cdots + n_k$.

Remark. The right-hand side of (9.5.6) is given by the Ewens sampling formula (9.2.11). Several explanations have been proposed for this result (Watterson (1989a), (1989b), Shimizu (1990)), but it seems to us that (9.5.9) below is the key observation.

Proof. Define $\sigma, \eta \in \mathcal{P}([0, 1]^n)$ by

$$(9.5.7) \quad \sigma(\cdot) = \int_{\mathcal{P}([0, 1]^n)} \mu(\cdot) \Pi(d\mu), \quad \eta(\cdot) = \int_{\mathcal{P}([0, 1])} \nu^n(\cdot) \Pi_0(d\nu).$$

Then

$$(9.5.8) \quad 0 = \int_{\mathcal{P}([0, 1]^n)} \mathcal{L}[\langle f, \mu \rangle] \Pi(d\mu) = \int_{\mathcal{P}([0, 1]^n)} \langle Af, \mu \rangle \Pi(d\mu) = \int_{[0, 1]^n} Af \, d\sigma;$$

here $\mathcal{L}[\langle f, \mu \rangle]$ is short for $(\mathcal{L}\varphi)(\mu)$, where $\varphi(\mu) = \langle f, \mu \rangle$. Moreover, letting \mathcal{L}_0 denote the generator of the Fleming-Viot process with type space $[0, 1]$ and mutation operator A_0 , we note that

$$(9.5.9) \quad \mathcal{L}_0[\langle f, \nu^n \rangle] = (2\lambda)^{-1} \langle Af, \nu^n \rangle$$

and hence

$$(9.5.10) \quad \begin{aligned} 0 &= \int_{\mathcal{P}([0, 1])} \mathcal{L}_0[\langle f, \nu^n \rangle] \Pi_0(d\nu) \\ &= (2\lambda)^{-1} \int_{\mathcal{P}([0, 1])} \langle Af, \nu^n \rangle \Pi_0(d\nu) = (2\lambda)^{-1} \int_{[0, 1]^n} Af \, d\eta. \end{aligned}$$

We conclude from (9.5.8), (9.5.10), and Lemma 9.5.1 that $\sigma = \eta$, and hence that (9.5.6) holds. \square

9.6. Infinitely-many-sites model without recombination. Here $E = [0, 1]^{\mathbf{Z}_+}$ and

$$(9.6.1) \quad (Af)(\mathbf{x}) = \frac{1}{2}\theta \int_0^1 (f(\xi, \mathbf{x}) - f(\mathbf{x})) \, d\xi.$$

The interpretation is as follows. The set of sites on the chromosome is identified with $[0, 1]$. An individual is of type $\mathbf{x} = (x_0, x_1, \dots) \in E$ if x_0, x_1, \dots is the sequence of sites at which mutations have occurred in the line of descent of that individual. For example, x_0 is the site at which the most recent mutation occurred to the individual or any of its ancestors.

The infinitely-many-sites model is due to Kimura (1969), (1971), who included recombination as well. Watterson (1975) studied the model without recombination, and the present formulation is due to Ethier and Griffiths (1987).

If $n \geq 1$, we say that $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in E^n$ is a *genealogical n -tree* if

- (a) the coordinates x_{ij} , $j \in \mathbf{Z}_+$, of \mathbf{x}_i are distinct for fixed $i \in \{1, \dots, n\}$;
- (b) whenever $i, i' \in \{1, \dots, n\}$, $j, j' \in \mathbf{Z}_+$, and $x_{ij} = x_{i'j'}$, we have $x_{i, j+l} = x_{i', j'+l}$ for all $l \geq 1$;
- (c) there exist $j_1, \dots, j_n \in \mathbf{Z}_+$ such that $x_{1j_1} = \cdots = x_{nj_n}$.

Informally, (a) means that mutations never occur more than once at the same site; (b) means that if two individuals have ancestors with the most recent mutations

in their respective lines of descent at the same site, then the ancestors are of the same type; and (c) means that every n individuals have a common ancestral type, hence a common ancestor.

For each $n \geq 1$, let $\mathcal{T}_n \subset E^n$ be the set of all genealogical n -trees, and let

$$(9.6.2) \quad \mathcal{P}_a^0(E) = \{\mu \in \mathcal{P}_a(E) : \mu^n(\mathcal{T}_n) = 1 \text{ for every } n \geq 1\}.$$

LEMMA 9.6.1. *The Fleming-Viot process with type space $E = [0, 1]^{\mathbb{Z}_+}$ and mutation operator A given by (9.6.1) has a unique stationary distribution Π , and $\Pi(\mathcal{P}_a^0(E)) = 1$. If $\{\mu_t, t \geq 0\}$ is a Fleming-Viot process with type space E and mutation operator A , and if $\mathbf{P}\{\mu_0 \in \mathcal{P}_a^0(E)\} = 1$, then $\mathbf{P}\{\mu_t \in \mathcal{P}_a^0(E) \text{ for all } t \geq 0\} = 1$.*

Given $n \geq 1$ and $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathcal{T}_n$, we say that $z \in [0, 1]$ is a *segregating site* of $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ if z appears in at least one but not all of the sequences $\mathbf{x}_1, \dots, \mathbf{x}_n$. Watterson (1975) determined the distribution of the number of segregating sites in a random sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ from μ , where μ is distributed according to Π . In the present setting, the result is due to Ethier and Griffiths (1987). For each $n \geq 1$ and $s \geq 0$, let $\mathcal{T}_{n,s} \subset E^n$ be the set of genealogical n -trees with s segregating sites.

THEOREM 9.6.2. *The stationary distribution Π of Lemma 9.6.1 satisfies*

$$(9.6.3) \quad \int_{\mathcal{P}(E)} \mu^n(\mathcal{T}_{n,s}) \Pi(d\mu) = \mathbf{P}\left\{\sum_{j=1}^{n-1} Y_j = s\right\},$$

for each $n \geq 2$ and $s \geq 0$, where Y_1, \dots, Y_{n-1} are independent with Y_j geometrically distributed on \mathbb{Z}_+ with parameter $j/(j + \theta)$.

Proof. For each $n \geq 1$ and $s \geq 0$, define $\varphi_{n,s} \in B(\mathcal{P}(E))$ by $\varphi_{n,s}(\mu) = \mu^n(\mathcal{T}_{n,s})$, and let $p_{n,s} = \int_{\mathcal{P}(E)} \varphi_{n,s}(\mu) \Pi(d\mu)$. Then, clearly $p_{1,s} = \delta_{s0}$, and the identity $\int_{\mathcal{P}(E)} (\mathcal{L}\varphi_{n,s})(\mu) \Pi(d\mu) = 0$ reduces (with some effort) to

$$(9.6.4) \quad \binom{n}{2} (p_{n-1,s} - p_{n,s}) + \frac{1}{2} n \theta (p_{n,s-1} - p_{n,s}) = 0, \quad n \geq 2, s \geq 0,$$

where $p_{n,-1} = 0$. Letting $q_n(z) = \sum_{s=0}^{\infty} p_{n,s} z^s$, we find from (9.6.4) that

$$(9.6.5) \quad (n-1)q_{n-1}(z) + \theta z q_n(z) = (n-1+\theta)q_n(z),$$

and hence

$$(9.6.6) \quad q_n(z) = \prod_{j=2}^n \frac{j-1}{j-1+\theta(1-z)},$$

as required. \square

9.7. Infinitely-many-neutral-alleles model with allelic genealogies.

Again, let S be a compact metric space. Here we define $E = S^{\mathbb{Z}_+} \times [0, \infty]^{\mathbb{Z}_+}$ and

$$(9.7.1) \quad (Af)(\mathbf{x}, \mathbf{a}) = \left(\frac{\partial}{\partial a_0} f\right)(\mathbf{x}, \mathbf{a}) + \frac{1}{2} \theta \int_S (f(\xi, \mathbf{x}, 0, \mathbf{a}) - f(\mathbf{x}, \mathbf{a})) P(x_0, d\xi),$$

where $\theta > 0$ and $P(x, d\xi)$ is a one-step Feller transition function on $S \times \mathcal{B}(S)$ satisfying (9.2.2); $\mathcal{D}(A)$ is defined in the obvious way (cf. (9.3.1)). An individual is of type $(\mathbf{x}, \mathbf{a}) = ((x_0, x_1, \dots), (a_0, a_1, \dots)) \in E$ if

x_0 = its allelic type,

a_0 = the age of x_0 ,

x_i = the allelic type of the individual that produced x_{i-1} by mutation, $i \geq 1$,

a_i = the age of x_i at the time x_{i-1} first appeared, $i \geq 1$.

This model is due to Ethier and Shiga (1993) and was motivated by the work of Takahata and Nei (1990) and Takahata (1991). It includes the processes of §§9.2, 9.3, and 9.6 as marginal processes. The stationary distribution is of primary interest.

LEMMA 9.7.1. *Assuming (9.2.2) and (9.3.4), the Fleming-Viot process with type space $E = S^{\mathbf{Z}_+} \times [0, \infty]^{\mathbf{Z}_+}$ and mutation operator A given by (9.7.1) has a unique stationary distribution $\Pi \in \mathcal{P}(E)$ and is weakly ergodic.*

As in §9.6, if $n \geq 1$, we say that $((\mathbf{x}_1, \mathbf{a}_1), \dots, (\mathbf{x}_n, \mathbf{a}_n)) \in E^n$ is a *genealogical n -tree* if

- (a) the coordinates x_{ij} , $j \in \mathbf{Z}_+$, of \mathbf{x}_i are distinct for fixed $i \in \{1, \dots, n\}$,
- (b) whenever $i, i' \in \{1, \dots, n\}$, $j, j' \in \mathbf{Z}_+$, and $x_{ij} = x_{i'j'}$, we have $(x_{i,j+l}, a_{i,j+l}) = (x_{i',j'+l}, a_{i',j'+l})$ for all $l \geq 1$ and $\sum_{l=0}^j a_{il} = \sum_{l=0}^{j'} a_{i'l}$,
- (c) there exist $j_1, \dots, j_n \in \mathbf{Z}_+$ such that $x_{1j_1} = \dots = x_{nj_n}$.

For each $n \geq 1$, let $\mathcal{T}_n \subset E^n$ denote the set of all genealogical n -trees, and define $\mathcal{P}_a^0(E)$ as in (9.6.2).

LEMMA 9.7.2. *Under the assumptions of Lemma 9.7.1, $\Pi(\mathcal{P}_a^0(E)) = 1$.*

We assume hereafter that

$$(9.7.2) \quad P(x, d\xi) \equiv \nu_0(d\xi),$$

where $\nu_0 \in \mathcal{P}(S)$ is nonatomic.

For $1 \leq k \leq n$, let $\pi(n, k)$ denote the set of partitions of $\{1, \dots, n\}$ into k nonempty subsets β_1, \dots, β_k , labeled so that $\min \beta_1 < \dots < \min \beta_k$. The *n -coalescent with mutation* (Kingman (1982a)) is a pure-jump Markov process with state space

$$(9.7.3) \quad \mathcal{E}_n(\mathbf{Z}_+) = \bigcup_{k=1}^n \{\pi(n, k) \times (\mathbf{Z}_+)^k\},$$

initial state $(\{1\}, \dots, \{n\}; 0, \dots, 0)$, and transitions from $(\beta_1, \dots, \beta_k; m_1, \dots, m_k)$ to

$$(9.7.4) \quad (\beta_1, \dots, \beta_{i-1}, \beta_i \cup \beta_j, \beta_{i+1}, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k; 0, \dots, 0)$$

with rate 1, $1 \leq i < j \leq k$, and to

$$(9.7.5) \quad (\beta_1, \dots, \beta_k; m_1, \dots, m_{i-1}, m_i + 1, m_{i+1}, \dots, m_k)$$

with rate $\theta/2$, $1 \leq i \leq k$. When there are k sets in the partition, the jump rate is $k(k-1+\theta)/2$. Transitions to (9.7.4) represent coalescences, whereas transitions to (9.7.5) represent mutations. Thus, m_1, \dots, m_k count the numbers of mutations in the k lines since the most recent coalescence.

For $i = 1, \dots, n$, let $M_i(t) = m_j(t)$ if $i \in \beta_j(t)$, $\tau_{i0} = \inf\{t > 0 : M_i(t) - M_i(t-) = 1\}$, and $\tau_{il} = \inf\{t > \tau_{i,l-1} : M_i(t) - M_i(t-) = 1\}$ for each $l \geq 1$. Arrange the distinct

elements of $\{\tau_{il} : i = 1, \dots, n, l \in \mathbf{Z}_+\}$ in ascending order $\tau_{(0)} < \tau_{(1)} < \tau_{(2)} < \dots$. Let $\xi_0, \xi_1, \xi_2, \dots$ be i.i.d. ν_0 . For $i = 1, \dots, n$ and $l \in \mathbf{Z}_+$, let $x_{il} = \xi_j$ if $\tau_{il} = \tau_{(j)}$.

THEOREM 9.7.3. *Assume (9.7.2) and let Π be as in Lemma 9.7.1. Then $((x_{i0}, x_{i1}, x_{i2}, \dots), (\tau_{i0}, \tau_{i1} - \tau_{i0}, \tau_{i2} - \tau_{i1}, \dots))$ ($i = 1, \dots, n$) have joint distribution $\int_{\mathcal{P}(E)} \mu^n(\cdot) \Pi(d\mu)$.*

This shows that the n th moment measure of Π (which can be interpreted as the joint distribution of a random sample of size n from μ , where μ is distributed according to Π) can be expressed in terms of the n -coalescent with mutation and ν_0 . The correspondence is not one-to-one, however. In particular, the coalescence times have no interpretation in the allelic genealogy. See Ethier and Shiga (1993) for the proof and further discussion.

REFERENCES

- J. F. CROW (1989), *Twenty-five years ago in genetics: the infinite allele model*, Genetics, 121, pp. 631–634.
- D. A. DAWSON (1978), *Geostochastic calculus*, Canadian J. Statist., 6, pp. 143–168.
- D. A. DAWSON AND K. J. HOCHBERG (1982), *Wandering random measures in the Fleming–Viot model*, Ann. Probab., 10, pp. 554–580.
- P. DONNELLY (1986), *Partition structures, Polya urns, the Ewens sampling formula and the ages of alleles*, Theoret. Popn. Biol., 30, pp. 271–288.
- P. DONNELLY AND P. JOYCE (1989), *Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex*, Stochastic Processes Appl., 31, pp. 89–103.
- (1991), *Consistent ordered sampling distributions: Characterization and convergence*, Adv. Appl. Probab., 23, pp. 229–258.
- P. DONNELLY AND T. G. KURTZ (1993), *A particle representation of infinite population genetic models*, manuscript.
- P. DONNELLY AND S. TAVARÉ (1986), *The ages of alleles and a coalescent*, Adv. Appl. Probab., 18, pp. 1–19; *Correction*, 18, p. 1023.
- (1987), *The population genealogy of the infinitely-many neutral alleles model*, J. Math. Biol., 25, pp. 381–391.
- R. DURRETT (1991), *Probability: Theory and Examples*, Wadsworth and Brooks/Cole, Pacific Grove, CA.
- E. B. DYNKIN (1989), *Three classes of infinite-dimensional diffusions*, J. Func. Anal., 86, pp. 75–110.
- S. N. ETHIER (1979), *A limit theorem for two-locus diffusion models in population genetics*, J. Appl. Probab., 16, pp. 402–408.
- (1981), *A class of infinite-dimensional diffusions occurring in population genetics*, Indiana Univ. Math. J., 30, pp. 925–935.
- (1990a), *The infinitely-many-neutral-alleles diffusion model with ages*, Adv. Appl. Probab., 22, pp. 1–24.
- (1990b), *The distribution of the frequencies of age-ordered alleles in a diffusion model*, Adv. Appl. Probab., 22, pp. 515–532.
- (1992a), *Equivalence of two descriptions of the ages of alleles*, J. Appl. Probab., 29, pp. 185–189.
- (1992b), *Eigenstructure of the infinitely-many-neutral-alleles diffusion model*, J. Appl. Probab., 29, pp. 487–498.
- S. N. ETHIER AND R. C. GRIFFITHS (1987), *The infinitely-many-sites model as a measure-valued diffusion*, Ann. Probab., 15, pp. 515–545.
- (1990a), *The neutral two-locus model as a measure-valued diffusion*, Adv. Appl. Probab., 22, pp. 773–786.
- (1990b), *On the two-locus sampling distribution*, J. Math. Biol., 29, pp. 131–159.
- (1993), *The transition function of a Fleming–Viot process*, Ann. Probab., to appear.
- S. N. ETHIER AND T. G. KURTZ (1981), *The infinitely-many-neutral-alleles diffusion model*, Adv. Appl. Probab., 13, pp. 429–452.
- (1986), *Markov Processes: Characterization and Convergence*, John Wiley, New York.

- S. N. ETHIER AND T. G. KURTZ (1987), *The infinitely-many-alleles model with selection as a measure-valued diffusion*, in *Stochastic Models in Biology*, M. Kimura, G. Kallianpur, and T. Hida, eds., *Lecture Notes in Biomathematics*, 70, Springer-Verlag, Berlin, pp. 72–86.
- (1992), *On the stationary distribution of the neutral diffusion model in population genetics*, *Ann. Appl. Probab.*, 2, pp. 24–35.
- (1993a), *Convergence to Fleming–Viot processes in the weak atomic topology*, *Stochastic Processes Appl.*, to appear.
- (1993b), *Coupling and ergodic theorems for Fleming–Viot processes*, manuscript.
- S. N. ETHIER AND T. NAGYLAKI (1980), *Diffusion approximations of Markov chains with two time scales and applications to population genetics*, *Adv. Appl. Probab.*, 12, pp. 14–49.
- S. N. ETHIER AND T. SHIGA (1993), *Neutral allelic genealogy*, manuscript.
- W. J. EWENS (1972), *The sampling theory of selectively neutral alleles*, *Theoret. Popn. Biol.*, 3, pp. 87–112.
- W. H. FLEMING AND M. VIOT (1979), *Some measure-valued Markov processes in population genetics theory*, *Indiana Univ. Math. J.*, 28, pp. 817–843.
- M. FUKUSHIMA AND D. W. STROOCK (1986), *Reversibility of solutions to martingale problems*, *Adv. Math. Suppl. Stud.*, 9, pp. 107–123.
- D. GRIFFEATH (1978), *Coupling methods for Markov processes*, in *Studies in Probability and Ergodic Theory*, G. Rota, ed., Academic Press, New York, pp. 1–43.
- R. C. GRIFFITHS (1979), *A transition density expansion for a multi-allele diffusion model*, *Adv. Appl. Probab.*, 11, pp. 310–325.
- (1981), *Neutral two-locus multiple allele models with recombination*, *Theoret. Popn. Biol.*, 19, pp. 169–186.
- R. C. GRIFFITHS AND G. A. WATTERSON (1990), *The number of alleles in multigene families*, *Theoret. Popn. Biol.*, 37, pp. 110–123.
- K. HANDA (1990), *A measure-valued diffusion process describing the stepping stone model with infinitely many alleles*, *Stochastic Processes Appl.*, 36, pp. 269–296.
- K. J. HOCHBERG (1991), *Measure-valued processes: techniques and applications*, in *Selected Proc. Sheffield Symp. Appl. Probab.*, IMS Lecture Notes–Monograph Series, 18, pp. 212–235.
- F. M. HOPPE (1987), *The sampling theory of neutral alleles and an urn model in population genetics*, *J. Math. Biol.*, 25, pp. 123–159.
- F. P. KELLY (1979), *Reversibility and Stochastic Networks*, John Wiley, Chichester.
- M. KIMURA (1969), *The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations*, *Genetics*, 61, pp. 893–903.
- (1971), *Theoretical foundations of population genetics at the molecular level*, *Theoret. Popn. Biol.*, 2, pp. 174–208.
- M. KIMURA AND J. F. CROW (1964), *The number of alleles that can be maintained in a finite population*, *Genetics*, 49, pp. 725–738.
- J. F. C. KINGMAN (1975), *Random discrete distributions*, *J. Roy. Statist. Soc. B*, 37, pp. 1–22.
- (1976), *Coherent random walks arising in some genetical models*, *Proc. Roy. Soc. London A*, 351, pp. 19–31.
- (1982a), *On the genealogy of large populations*, *J. Appl. Probab.*, 19A, pp. 27–43.
- (1982b), *The coalescent*, *Stochastic Processes Appl.*, 13, pp. 235–248.
- N. KONNO AND T. SHIGA (1988), *Stochastic partial differential equations for some measure-valued diffusions*, *Probab. Th. Rel. Fields*, 79, pp. 201–225.
- T. G. KURTZ (1981), *Approximation of Population Processes*, CBMS–NSF Regional Conf. Ser. Appl. Math., Vol. 36, Society for Industrial and Applied Mathematics, Philadelphia, PA.
- R. A. LITTLER (1972), *Multidimensional Stochastic Models in Genetics*, Ph.D. thesis, Monash University.
- P. A. P. MORAN (1958), *Random processes in genetics*, *Proc. Camb. Philos. Soc.*, 54, pp. 60–71.
- (1975), *Wandering distributions and the electrophoretic profile*, *Theoret. Popn. Biol.*, 8, pp. 318–330.
- (1976), *Wandering distributions and the electrophoretic profile*, II, *Theoret. Popn. Biol.*, 10, pp. 145–149.
- T. NAGYLAKI (1990), *Models and approximations for random genetic drift*, *Theoret. Popn. Biol.*, 37, pp. 192–212.
- T. OHTA (1982), *Allelic and nonallelic homology of a supergene family*, *Proc. Natl. Acad. Sci. USA*, 79, pp. 3251–3254.
- (1983), *On the evolution of multigene families*, *Theoret. Popn. Biol.*, 23, pp. 216–240.

- T. OHTA AND M. KIMURA (1973), *A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population*, Genet. Res. Camb., 22, pp. 201–204.
- G. P. PATIL AND C. TAILLIE (1977), *Diversity as a concept and its implications for random communities*, Bull. Internat. Statist. Inst., 47, pp. 497–515.
- M. REIMERS (1989), *One dimensional stochastic partial differential equations and the branching measure diffusion*, Probab. Th. Rel. Fields, 81, pp. 319–340.
- (1993), *A new result on the support of the Fleming–Viot process, proved by non-standard construction*, manuscript.
- S. SAWYER (1977), *On the past history of an allele now known to have frequency p* , J. Appl. Probab., 14, pp. 439–450.
- T. SHIGA (1982), *Wandering phenomena in infinite allelic diffusion models*, Adv. Appl. Probab., 14, pp. 457–483.
- (1990), *A stochastic equation based on a Poisson system for a class of measure-valued diffusion processes*, J. Math. Kyoto Univ., 30, pp. 245–279.
- N. SHIMAKURA (1977), *Equations différentielles provenant de la génétique des populations*, Tôhoku Math. J., 29, pp. 287–318.
- (1981), *Formulas for diffusion approximations of some gene frequency models*, J. Math. Kyoto Univ., 21, pp. 19–45.
- A. SHIMIZU (1985), *Diffusion approximation of an infinite allele model incorporating gene conversion*, in Population Genetics and Molecular Evolution, T. Ohta and K. Aoki, eds., Japan Sci. Soc. Press, Tokyo, Springer-Verlag, Berlin, pp. 243–255.
- (1987), *Stationary distribution of a diffusion process taking values in probability distributions on the partitions*, in Stochastic Models in Biology, M. Kimura, G. Kallianpur, and T. Hida, eds., Lecture Notes in Biomathematics, 70, Springer-Verlag, Berlin, pp. 100–114.
- (1990), *A measure valued diffusion process describing an n locus model incorporating gene conversion*, Nagoya Math. J., 119, pp. 81–92.
- C. STROBECK AND K. MORGAN (1978), *The effect of intragenic recombination on the number of alleles in a finite population*, Genetics, 88, pp. 829–844.
- N. TAKAHATA (1991), *A trend in population genetics theory*, in New Aspects of the Genetics of Molecular Evolution, M. Kimura and N. Takahata, eds., Japan Sci. Soc. Press, Tokyo, Springer-Verlag, Berlin, pp. 27–47.
- N. TAKAHATA AND M. NEI (1990), *Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci*, Genetics, 124, pp. 967–978.
- S. TAVARÉ (1984), *Line-of-descent and genealogical processes, and their applications in population genetics models*, Theoret. Popn. Biol., 26, pp. 119–164.
- J. VAILLANCOURT (1990a), *Interacting Fleming–Viot processes*, Stochastic Processes Appl., 36, pp. 45–57.
- (1990b), *On the scaling theorem for interacting Fleming–Viot processes*, Stochastic Processes Appl., 36, pp. 263–267.
- G. A. WATTERSON (1975), *On the number of segregating sites in genetical models without recombination*, Theoret. Popn. Biol., 7, pp. 256–276.
- (1976), *The stationary distribution of the infinitely-many neutral alleles diffusion model*, J. Appl. Probab., 13, pp. 639–651.
- (1989a), *Allele frequencies in multigene families. I. Diffusion equation approach*, Theoret. Popn. Biol., 35, pp. 142–160.
- (1989b), *Allele frequencies in multigene families. II. Coalescent approach*, Theoret. Popn. Biol., 35, pp. 161–180.
- G. A. WATTERSON AND H. A. GUESS (1977), *Is the most frequent allele the oldest?* Theoret. Popn. Biol., 11, pp. 141–160.
- S. WRIGHT (1949), *Adaptation and selection*, in Genetics, Paleontology, and Evolution, G. L. Jepson, E. Mayr, and G. G. Simpson, eds., Princeton University Press, Princeton, NJ, pp. 365–389.

CURVATURE-DRIVEN FLOWS: A VARIATIONAL APPROACH*

FRED ALMGREN†, JEAN E. TAYLOR‡, AND LIHE WANG§

This paper is dedicated to Wendell Fleming on the occasion of his 65th birthday.

Abstract. This paper introduces a new mathematical approach to the study of time evolutions of solids $K(t)$ in n -space whose boundaries move with velocity equal to the weighted mean curvature derived from the boundary surface energy $\Phi(\partial K(t))$. These “flat Φ curvature flows” are limits of sequences of solutions to variational problems in which a sum of surface and bulk energy is minimized. The construction works equally well for smooth elliptic Φ ’s, for nondifferentiable crystalline Φ ’s, and for anything in between. The flows agree with classical smooth flows when the data is smooth and elliptic in any dimension and coincide with motion by crystalline curvature for polyhedral curves in the plane.

Key words. curvature evolution, flat curvature flow, motion-by-mean curvature, weighted mean curvature, currents, calculus of variations, geometric measure theory

AMS(MOS) subject classifications. 35A15, 35K99, 49F22, 58E99

1. Introduction. This paper is a contribution to the study of time evolutions of solids $K(t)$ in space whose boundaries move with velocity equal to the weighted mean curvature derived from the boundary surface energy $\Phi(\partial K(t))$. By “weighted mean curvature” at boundary point p , we mean (intuitively) the rate at p at which surface energy decreases with unit rate of volume change. If $\partial K(t)$ is a smooth hypersurface and if the surface energy $\Phi(\partial K(t))$ is the surface area of $\partial K(t)$, such evolution is called “motion-by-mean curvature.” For other smooth surface energies, weighted mean curvature can be expressed as a weighted sum of principal curvatures. In this paper, we give a reasonable construction of such motion for general convex parametric surface energy functions Φ . Our estimates apply, in particular, to crystalline surface energies, which are not differentiable. (The paper [BSS] by Barles, Soner, and Souganidis reviews and unifies several approaches to a class of geometric flows including motion by mean curvature. The paper [TCH] of Taylor, Cahn, and Handwerker gives a general discussion of geometric models of crystal growth, including both metallurgical applications and even more formulations of motion by weighted mean curvature; it includes references to contributions by S. Angenent, K. Brakke, L. Bronsard, Y.-G. Chen, K. Ecker, L. C. Evans, M. Gage, Y. Giga, S. Goto, M. Grayson, R. Hamilton, G. Huisken, T. Ilmanen, R. V. Kohn, W. Mullins, S. Osher, J. A. Sethian, H. M. Soner, P. E. Souganidis, J. Spruck, J. E. Taylor, and many others.)

* Received by the editors June 1, 1992; accepted for publication May 20, 1992.

† Department of Mathematics, Princeton University, Princeton, New Jersey 08544. This author is a permanent faculty member of The National Science and Technology Research Center for Computation and Visualization of Geometric Structures (The Geometry Center), Minneapolis, Minnesota 55415. During the preparation of this paper, this author was also supported in part by grants from the National Science Foundation and was a visitor at the Centre for Mathematical Analysis of the Australian National University, Canberra, Australia and at the Aspen Center for Physics, Aspen, Colorado 81611. During the preparation of this paper, this author was a visiting member of the Institute for Advanced Study, Princeton, New Jersey, 08540.

‡ Department of Mathematics, Rutgers University, New Brunswick, New Jersey, 08903. This author is a permanent faculty member of The National Science and Technology Research Center for Computation and Visualization of Geometric Structures (The Geometry Center), Minneapolis, Minnesota 55415. During the preparation of this paper, this author was also supported in part by grants from the National Science Foundation and was a visitor at the Centre for Mathematical Analysis of the Australian National University, Canberra, Australia and at the Aspen Center for Physics, Aspen, Colorado 81611.

§ Department of Mathematics, Princeton University, Princeton, New Jersey 08544.

For a given initial solid $K(0)$ and parametric surface energy function Φ , we select a sequence of solid positions $K_j(t)$ associated with timesteps of size Δt_j . Compactness theorems for integral currents, invoked in Cantor's diagonal process, guarantee the existence of a subsequence $j(1), j(2), j(3), \dots$ of $1, 2, 3, \dots$ such that, for a dense set of times, the currents $[K_{j(i)}(t)]$ will converge to a limit current $[K(t)]$ as $i \rightarrow \infty$; this fact, in itself, does not preclude the possibility that the $[K(t)]$'s might vary in a wildly discontinuous way. As our pivotal result, we prove a priori estimates for the discrete evolutions K_j , which imply the Hölder continuity of the limit K in the appropriate metric (see § 4). We also show that, if the initial boundary lies outside a given Wulff shape, then the boundary later lies outside a suitably scaled Wulff shape, and, if the initial solid lies inside a given Wulff shape, then, it later lies inside a suitably scaled Wulff shape (see § 5). Additionally, we establish sufficient conditions, so that the evolution be smooth locally in space-time (see § 6). We further show (see § 7) that, if $\partial K(0)$ is a smooth hypersurface and Φ is smooth and elliptic, then there is a smooth Φ curvature flow beginning with $\partial K(0)$ and continuing for at least a short while. For smooth and elliptic Φ 's, flat Φ curvature flows coincide with smooth flows for as long as the latter remain smooth. In § 8 we give several examples.

In a related paper [AT], Almgren and Taylor show that, for polyhedral curves in the plane driven by a crystalline even Φ , the motion constructed in this paper typically coincides with motion by crystalline curvature produced by direct integration of ordinary differential equations.

The mathematics contained in this paper largely belongs to the field of geometric measure theory. Wendell Fleming made fundamental and pivotal contributions to the field during the 1950s and 1960s, part of which time the first author was a graduate student in mathematics at Brown University. (The 1960 paper [FF] laid many of the foundations of modern geometric measure theory; this paper was awarded the 1986 Steele Prize of the American Mathematical Society for a seminal contribution to Mathematics.)

2. Questions and answers about Φ curvature flows.

2.1. What are the ingredients of Φ curvature flows? The context in which we study curvature flows is the following.

2.1.1. Environment. The ambient space in which we study curvature flows is n -dimensional Euclidean space \mathbb{R}^n , which is measured by n -dimensional Lebesgue measure \mathcal{L}^n and $(n-1)$ -dimensional Hausdorff measure \mathcal{H}^{n-1} . When we write $A \subset_{n-1} B$, we mean $\mathcal{H}^{n-1}(A \sim B) = 0$. We set

$$\mathbb{B}^n(p, r) = \mathbb{R}^n \cap \{x: |x - p| \leq r\}, \quad \alpha(n) = \mathcal{L}^n(\mathbb{B}^n(0, 1)).$$

We denote by $\beta(n)$ the covering multiplicity constant of the Besicovitch covering theorem (denoted $2\zeta + 1$ in [F, § 2.8.14]). We additionally denote by \mathcal{C} the collection of all compact subsets of \mathbb{R}^n equipped with the Hausdorff distance function \mathbb{HD} , so that

$$\mathbb{HD}(A, B) = \sup \{ \inf \{|a - b|: a \in A\}: b \in B \} + \sup \{ \inf \{|a - b|: b \in B\}: a \in A \}$$

for $A, B \in \mathcal{C}$. There seem to be no particular obstacles to the study of curvature motions in other ambient manifolds.

2.1.2. Solids and their boundaries. By *solids*, we mean n -dimensional integral currents $[K] = \mathbb{E}^n \llcorner K$ in \mathbb{R}^n associated with bounded \mathcal{L}^n measurable subsets K of \mathbb{R}^n having finite perimeter. Currents seem the natural setting for our study, since, in particular, sets differing in measure zero are identified; various definitions and properties of currents are set forth in § 3.1, below. We denote by \mathcal{K} the space of such solids,

and, when we write $[K] \in \mathcal{K}$, we mean, in particular, that K is such a set of finite perimeter. Associated with each such K is the rectifiable set ∂K of points p in \mathbb{R}^n at which K has a well-defined unit exterior normal vector $\mathbf{n}_K(p)$. As currents, we can write $\partial[K] = \mathcal{H}^{n-1} \llcorner \partial K \wedge * \mathbf{n}_K$. We denote by $\partial\mathcal{K}$ the space of $(n-1)$ -dimensional integral currents $\partial[K]$ associated with $[K]$ s in \mathcal{K} . The metric we use on solids is given by the mass norm \mathbf{M} , which in this case can be written as

$$\mathbf{M}([K] - [L]) = \mathcal{L}^n(K \triangle L);$$

here

$$K \triangle L = ((K \sim L) \cup (L \sim K)) = K \cap (\mathbb{R}^n \sim L) \cup (\mathbb{R}^n \sim K) \cap L$$

denotes the symmetric difference between K and L . There is a corresponding \mathbb{G} norm on boundaries given by setting

$$\mathbb{G}(\partial[K] - \partial[L]) = \mathbf{M}([K] - [L]).$$

The norm \mathbb{G} is essentially equivalent to the flat norm on currents. The support $\text{spt } \partial[K]$ of each member $\partial[K]$ of $\partial\mathcal{K}$, of course, is a set belonging to \mathcal{C} .

2.1.3. Surface energy integrands and their integrals. We denote by

$$\Phi: \mathbb{R}^n \rightarrow \mathbb{R}^+$$

a fixed *parametric surface energy integrand*; this means that Φ is continuous and is positively homogeneous of degree one (i.e., $\Phi(\lambda v) = \lambda \Phi(v)$ if λ is nonnegative). We also assume that Φ is positive on \mathbb{R}^n except at the origin and that Φ is convex. (This assumption of convexity is not necessarily true for the surface energy functions of physical crystalline materials. When energies are not convex, it is customary to replace the actual energy by the smallest convex energy that is not less than the actual energy and then to understand surfaces to have tangent planes in the sense of varifold geometry. The Wulff shape of the convexified energy (see § 5) is the same as that of the original energy.) We call Φ *even*, provided that $\Phi(-v) = \Phi(v)$ for each v . We call Φ *smooth*, provided that the function Φ is three times Hölder continuously differentiable except at the origin. We call a smooth Φ *elliptic* provided that Φ is uniformly convex in the sense that the restriction of Φ to any arc length parametrized line in \mathbb{R}^n not containing the origin has strictly positive second derivatives. We further denote

$$0 < \Phi_0 =: \inf \{ \Phi(v) : |v| = 1 \} \leq \Phi^0 =: \sup \{ \Phi(v) : |v| = 1 \} < \infty$$

as lower and upper bounds for surface energy density on unit normal vectors. By the integral of Φ over $\partial[K]$, we mean the number

$$\Phi(\partial[K]) = \int_{p \in \partial K} \Phi(\mathbf{n}_K(p)) \, d\mathcal{H}^{n-1} p.$$

Similarly,

$$\Phi(-\partial[K]) = \int_{p \in \partial K} \Phi(-\mathbf{n}_K(p)) \, d\mathcal{H}^{n-1} p,$$

since $-\partial[K]$ is the boundary of $-[K]$, the current with negative orientation associated to K . The *mass integrand* \mathbf{M} is given by setting $\mathbf{M}(v) = |v|$ for each v , so that $\mathbf{M}(\partial[K]) = \mathcal{H}^{n-1}(\partial K)$.

2.2. In the smooth case, what is Φ curvature, and what is its relation with Φ ? Suppose S is a twice continuously differentiable closed submanifold of \mathbb{R}^n oriented by unit normal vector field \mathbf{n} . Associated with S is the smooth normal Φ curvature vector field

$$H = H_{\Phi, S}: S \rightarrow \mathbb{R}^n$$

given by setting for each $p \in S$,

$$H_{\Phi, S}(p) = \mathbf{n}(p) \operatorname{trace} (D^2\Phi(\mathbf{n}(p)) \circ D^2R(p)).$$

Here R is any twice continuously differentiable real-valued function defined in some neighborhood U of p in \mathbb{R}^n for which

$$S \cap U = R^{-1}\{0\} \quad \text{and} \quad \nabla R(p) = -\mathbf{n}(p).$$

Also, $D^2R(p)$ here denotes the $n \times n$ Hessian matrix of second derivatives of R evaluated at p , and $D^2\Phi(\mathbf{n}_K(p))$ denotes the $n \times n$ Hessian matrix of second derivatives of Φ evaluated at the unit normal vector $\mathbf{n}(p)$; R is defined on the ambient \mathbb{R}^n , while Φ is defined on the \mathbb{R}^n of normal vectors. (If S is the sphere of radius r in \mathbb{R}^n oriented as the boundary of a ball and $\Phi(v) = \mathbf{M}(v) = |v|$ (so that $\Phi(S) = \mathcal{H}^{n-1}(S)$ is the surface area of S), then we could take $R(x) = (2r)^{-1}(r^2 - |x|^2)$ and compute $H_{\Phi, S}(p) = -(n-1)p/r$ for each p in the sphere of radius r .)

If $G_t: \mathbb{R}^n \rightarrow \mathbb{R}^n$ ($t \in \mathbb{R}$) is a deformation of \mathbb{R}^n starting with the identity at time $t=0$ and having initial velocity vector field $g = (d/dt)G_t|_{t=0}$ (which vanishes on ∂S), then

$$\left. \frac{d}{dt} \Phi(G_t(S)) \right|_{t=0} = - \int_{p \in S} g(p) \cdot H_{\Phi, S}(p) d\mathcal{H}^{n-1}p.$$

If $S = \partial K$ for some $[K] \in \mathcal{K}$ and $\mathbf{n} = \mathbf{n}_K$, then we write $H = H_{\Phi, \partial K}$ and have, correspondingly,

$$\left. \frac{d}{dt} \Phi(G_t(\partial K)) \right|_{t=0} = - \int_{p \in \partial K} g(p) \cdot H_{\Phi, \partial K}(p) d\mathcal{H}^{n-1}p.$$

Since

$$\left. \frac{d}{dt} \mathcal{L}^n(G_t(K)) \right|_{t=0} = \int_{p \in \partial K} g(p) \cdot \mathbf{n}_K(p) d\mathcal{H}^{n-1}p,$$

we can intuitively regard $H_{\Phi, \partial K}(p)$ as the infinitesimal rate of change of surface energy with volume at p .

2.3. What is smooth Φ curvature flow? Suppose that Φ is smooth and that

$$[K(\cdot)]: (a, b) \rightarrow \mathcal{K}$$

is a time parametrized family of solids. We say that the $\partial[K(t)]$ s constitute a *smooth Φ curvature flow*, provided that the $\partial K(t)$ s are a smoothly varying family of hypersurfaces and that, for each $a < s < b$ and each $p \in \partial K(s)$, the normal velocity of the $\partial K(t)$ s at p at time s equals $H_{\Phi, \partial K(s)}(p)$. Obvious definitions apply for flows on closed and half open intervals of time. Also, the notion of smooth Φ curvature flows has an obvious localization in space.

2.4. Do smooth Φ curvature flows exist? If Φ is elliptic and $\partial K(0)$ is a three times Hölder continuously differentiable submanifold, then there will exist a smooth Φ curvature flow $\partial[K(t)]$ beginning with $\partial[K(0)]$ and defined for some interval of times t of positive length; see Theorem 7.1. Examples show that such flows can develop singularities in finite time.

2.5. What are nonparametric smooth Φ curvature flows? Suppose that Φ is elliptic. Suppose also that U is an open subset of $\mathbb{R}^{n-1} \times \mathbb{R}^+$ and $f: U \rightarrow \mathbb{R}$ is twice continuously differentiable. We say that f is a *nonparametric smooth Φ curvature flow*, provided that the hypersurfaces $\Sigma(t) = \mathbb{R}^n \cap \{(x, f(x, t)): (x, t) \in U\}$ are varying smoothly with normal velocities equal to $H_{\Phi, \Sigma(t)}$ for some continuous choice of normal vectors.

2.6. What is a flat Φ curvature flow? By a *flat Φ curvature flow*, we mean any function $\partial[K(\cdot)]: \mathbb{R}^+ \rightarrow \partial\mathcal{K}$ that is constructed by the following variational procedure. Whenever $[K] \in \mathcal{K}$ and $[L] \in \mathcal{K}$ and Δt is a positive number, we set

$$\mathcal{E}([K], [L], \Delta t) = \Phi(\partial[L]) + \frac{1}{\Delta t} \int_{p \in K \triangle L} \text{dist}(p, \partial K) d\mathcal{L}^n p.$$

Suppose then that $[K(0)] \in \mathcal{K}$ is a given initial solid position. For each fixed integer j , we set $\Delta t = 2^{-j}$ and choose

$$[K_j(\cdot)]: \mathbb{R}^+ \rightarrow \mathcal{K}$$

by the inductive requirements that

$$[K_j(0)] = [K(0)],$$

and, for each $k = 0, 1, 2, 3, \dots$, $[K_j(k\Delta t + \Delta t)]$ is chosen so that

$$\mathcal{E}([K_j(k\Delta t)], [K_j(k\Delta t + \Delta t)], \Delta t) = \mathcal{E}([K_j(k\Delta t)], [L], \Delta t)$$

and

$$K_j(k\Delta t + s) = K_j(k\Delta t + \Delta t)$$

for each $0 < s \leq \Delta t$. This defines *approximate flows* $\partial[K_j(t)]$ for $j = 1, 2, 3, \dots$ and all $t \geq 0$. A function $\partial[K(\cdot)]: \mathbb{R}^+ \rightarrow \partial\mathcal{K}$ is called a *flat Φ curvature flow*, provided that

$$\lim_{i \rightarrow \infty} \mathbb{G}(\partial[K(t)] - \partial[K_{j(i)}(t)]) = 0$$

locally uniformly in time t for some approximate flows $\partial[K_j(t)]$ and some subsequence $j(1), j(2), j(3), \dots$ of $1, 2, 3, \dots$.

2.7. Do flat Φ curvature flows exist? We show in § 4.4, below, that there are a priori \mathbb{G} Hölder continuity estimates for the approximate flows constructed above. These estimates imply the existence of the required convergent subsequences in § 4.5.

2.8. Are flat Φ curvature flows smooth? Suppose that Φ is smooth and elliptic and that the boundary of our initial solid is a three times Hölder continuously differentiable hypersurface. Then, at least for a short time, a smooth Φ curvature flow will exist, beginning with this hypersurface; see Theorem 7.1. For as long as this smooth flow does exist, it will coincide with any flat Φ curvature flow; see Theorem 7.4.

2.9. What are viscosity Φ curvature flows? Suppose that Φ is smooth and that $A: (a, b) \rightarrow \mathcal{C}$ is continuous in the $\mathbb{H}\mathbb{D}$ topology so that $\Sigma = \mathbb{R}^n \times (a, b) \cap \{(x, t): x \in A(t)\}$ is closed in $\mathbb{R}^n \times (a, b)$. Assume also that $\sigma: \mathbb{R}^n \times (a, b) \rightarrow \Sigma \rightarrow \{-1, 1\}$ is a continuous orientation function; σ is supposed to take value -1 in the crystal, and $+1$ otherwise. We say that A is a *viscosity Φ curvature flow* (with respect to σ), provided that the following condition holds: Suppose that $g: \mathbb{R}^n \times (a, b) \rightarrow \mathbb{R}$ is any twice continuously differentiable test function so that $\Gamma = \mathbb{R}^n \times (a, b) \cap \{(x, t): g(x, t) = 0\}$ is closed and $\Gamma^+ = \mathbb{R}^n \times (a, b) \cap \{(x, t): g(x, t) > 0\}$ is open. In the case where $\Sigma \cap \Gamma^+$ is empty and $\Sigma \cap \Gamma$ consists of a single point (p, s) at which $\nabla_x g(p, s) \neq 0$, we then require that

$$\frac{\partial g}{\partial t}(p, s) \geq \text{trace} \left(D^2 \Phi \left(\pm \frac{\nabla_x g}{|\nabla_x g|}(ps) \right) \circ D_x^2 g(p, s) \right);$$

here we take the $+$ sign if σ is positive in Γ_+ near (p, s) , and the $-$ sign otherwise. It is easy to check that an equivalent definition results if we replace the condition that $\Sigma \cap \Gamma$ consist of a single point (p, s) by the requirement that $(p, s) \in \Sigma \cap \Gamma$.

2.10. Do viscosity Φ curvature flows exist? Viscosity Φ curvature flows (as defined above) for smooth elliptic Φ are closely related to the viscosity solutions of the level set approach, which were initially shown to exist and be unique for any continuous initial data by [CGG] (for $\Phi = \mathbf{M}$, this was simultaneously shown by [ES1]). The current status of that and related approaches is given by [BSS]. In any of those contexts or in that of § 2.9 above, any smooth Φ curvature flow can be shown to be a viscosity Φ curvature flow. Now, suppose that $\partial[K(t)]$ is a flat Φ curvature flow constructed from approximating flows $\partial[K_j(t)]$. Let $A(t) = \text{spt } \partial[K(t)]$ and $A_j(t) = \text{spt } \partial[K_j(t)]$ for times t . In case $A(t)$ is continuous in the $\mathbb{H}\mathbb{D}$ topology for $a < t < b$ and $\mathbb{H}\mathbb{D}(A_j(t), A(t)) \rightarrow 0$ locally uniformly in t for $a < t < b$, then $A(t)$ is a viscosity Φ curvature flow (in § 2.9), as we show in § 6.2.

2.11. Are viscosity Φ curvature flows smooth? The most general information we know about this question is the following. Suppose that U is an open subset of $\mathbb{R}^{n-1} \times \mathbb{R}^+$ and that $f: U \rightarrow \mathbb{R}$ is a Lipschitz function. A necessary and sufficient condition that f be a nonparametric smooth Φ curvature flow is that the function $t \mapsto \{(x, f(x, t))\}$ be a viscosity Φ curvature flow. The Hölder continuous differentiability of f follows from the work [CW] of Caffarelli and Wang, while the twice Hölder continuous differentiability follows from work [W] of Wang.

2.12. Why is it reasonable to think of flat Φ curvature flows as curvature flows? Suppose that Φ is elliptic. We show in §§ 3.5–3.8 that the general surface regularity theory of [Bo] (or of [A1] in case Φ is even), together with higher differentiability estimates from the associated Euler-Lagrange partial differential equations (PDEs) guarantee that each $\partial K_j(t)$ ($t > \Delta t$) will \mathcal{H}^{n-1} almost everywhere be a smooth submanifold of \mathbb{R}^n and that, at each regular point p of $\partial K_j(t)$,

$$H_{\Phi, \partial K_j(t)}(p) = \frac{\text{dist}(p, \partial K_j(t - \Delta t))}{\Delta t} \mathbf{n}(p);$$

here $\Delta t = 2^{-j}$ is the timestep associated with the approximations $K_j(t)$, and $\mathbf{n}(p) = \mathbf{n}_\Omega(p)$ is the unit exterior normal vector to the symmetric difference $\Omega = K_j(t - \Delta t) \triangle K_j(t)$. In this sense, at each stage of the approximation, we have a reasonable approximation to a curvature flow if we accept

$$\frac{\text{dist}(p, \partial K_j(t - \Delta t))}{\Delta t}$$

as a reasonable approximation of interface speed and are willing to measure curvature at p in $\partial K_j(t)$ rather than, say, somewhere on $\partial K_j(t - \Delta t)$.

For smooth elliptic Φ and smooth initial data, flat Φ curvature flow coincides with smooth Φ curvature flow at least until singularities develop in the smooth flow, as shown in Theorem 7.4.

For polyhedral curves in the plane and a crystalline Φ , which is also even, flat Φ curvature flow typically coincides with motion by crystalline curvature computed by integration of ordinary differential equations, as shown in [AT].

For all other surface energy functions, the construction also appears to be a reasonable one and has strong a priori estimates.

2.13. What are reasons for interest in general flat Φ curvature flows? Our original interest in flat flows arose in the study of the changing geometry of a crystal as it evolves within its melt in a cold environment either by melting crystal or freezing melt [AW]; our interest is both theoretical and computational. In one model of such evolution, we intuitively consider the equation

$$v = M(\Omega + \text{wmc})$$

in which v denotes the surface normal velocity, Ω is the undercooling below the freezing temperature T_0 of a planar interface, wmc is the weighted mean curvature of the interface measuring the desire of the interface to decrease its surface energy, and M is the mobility of the interface measuring the response of the interface to the driving pressures of undercooling and surface tension. This model also applies to a wide range of other interface motion problems, involving chemical- as well as temperature-dependent bulk driving forces. In the freezing problem, we usually add a heat of fusion as new regions crystallize (or take heat away if melting occurs). This raises the temperature and slows the growth (since the amount of undercooling is decreased). One of the limiting factors in the rate of growth is thus the speed at which this added heat can diffuse away to the cold environment. The more subtle part of the problem, however, involves the curvature wmc , since this changes with changes in crystal shape. To focus on this difficulty, we can assume that the ambient temperature is T_0 and that heat diffuses infinitely fast; under this assumption, Ω is always zero. There seems little general theoretical or experimental knowledge about the form of the mobility function M (only its sign seems determined by classical thermodynamics [Gu]). To focus on the curvature difficulty, we also simply assume that the mobility M is identically equal to one. We are thus led to the study of our heuristic equation that

$$v_{\partial K(t)} = H_{\Phi, \partial K(t)},$$

to which we referred above. In this case, the only relevant driving force is that seeking to reduce surface energy so that any initial crystal will ultimately melt. By doing this, we are able to concentrate on the following central difficulties involved in making theoretical and computational sense of such flows:

(i) It is necessarily the case (by example) that such flows will develop singularities in finite time (such as necks pinching off or separate crystals shrinking to points at different times). Since we wish the flows to be defined for all time, it is necessary to find a construction or definition that enables us to continue the evolution even in the presence of singularities (including unbounded or undefined curvatures);

(ii) Generally, it is numerically difficult and unreliable to try to obtain curvature information from combinatorial representations of surfaces [U], [DS]. (Note that the crystalline curvature approach avoids this problem by using a natural parametrization derived from the Gauss map; see [T2] for the case of the motion of polyhedral curves in the plane);

(iii) Many naturally occurring surface energy functions for crystalline materials are not smooth and elliptic. Surface energies play a demonstrably important role in the geometry of the crystallization process for such surface energy functions, but their influence is not expressible in naively formulated notions of curvature, since the rate of change of surface energy with volume for faceted shapes or shapes with corners typically is either zero or infinite. We are thus led to find a process in which surface energy wants to be minimized but is constrained by a cost of adding or deleting from the solid region.

2.14. Is it possible to compute flat flows? Generally, it is difficult to perform the variational minimizations required by our procedure, but, when $n = 2$ or the variational problem can be legitimately confined to a finite-dimensional subspace, there has been substantial progress, including working codes and graphics. For crystalline surface energies, this reduction to a low-dimensional subspace has been shown to hold for $n = 2$ [AT], and the appropriate approximations for $n = 3$ seem to be in hand. These situations are illustrated in papers and videotapes by Taylor [T2]–[T4]. Also, for $n = 2$, computations including both curvature and bulk driving forces involving quantities that diffuse are illustrated in work of Roosen and Taylor [RT], [R]. In a related paper of Almgren [Al], the mobility M is taken to be infinite so that bulk driving force balances weighted mean curvature.

2.15. What are the main missing ingredients about Φ curvature flows in the flat sense? Perhaps the biggest missing estimate is one that would guarantee that $\mathcal{H}^{n-1}(\text{spt } \partial[K(t)] \sim \partial K(t)) = 0$ for times t . This does hold in the approximations to the flow as shown in § 3.4. However, (except when $\Phi = \mathbf{M}$) we know no way to preclude the possibility that approximations develop filigree structures that become increasingly elaborate as $\Delta t \downarrow 0$; if this does happen, it might well lead to limit currents with excessively large supports. Another missing estimate in the smooth case is one that would guarantee that, if $\partial K(t)$ were measure theoretically nearly a flat disk somewhere, then, a short time later, it would become a smooth nearly flat disk evolving smoothly. A third missing estimate is one that would guarantee that combining a flat flow from $t = 0$ to $t = T_1$ with one from $t = T_1$ to $t = T_2$ would yield a flat flow from $t = 0$ to $t = T_2$.

3. Existence and structure of \mathcal{E} minimizers. This section establishes existence and properties of \mathcal{E} minimizers including regularity in the case where surface energy is smooth and elliptic. We first set forth basic properties of currents needed in our analysis. The existence of \mathcal{E} minimizers is then a consequence of a compactness theorem for integral currents in the \mathbb{G} topology. We establish various properties of \mathcal{E} minimizers and show their relation to the (Φ, ω, δ) minimal currents of Bombieri and the (γ, δ) restricted sets and the $(\Phi, \varepsilon, \delta)$ minimal sets of Almgren; surface energy integrals over sets are defined only for even Φ . If our surface energy function Φ is smooth and elliptic, such minimality alone implies that the supports of \mathcal{E} minimizers are almost everywhere continuously differentiable submanifolds of \mathbb{R}^n . The particular form of our \mathcal{E} minimization, however, enables us to conclude higher differentiability. In particular, if K is a minimizer in $\mathcal{E}([K_0], [K], \Delta t)$ and ∂K_0 is a $C^{k,\alpha}$ submanifold, then those parts of ∂K near ∂K_0 will be a C^{k+2} submanifold almost everywhere. If $n = 2$ or 3 , there are no singularities in \mathcal{E} minimizers for elliptic Φ .

3.1. Integral currents and their properties. The present study of flat Φ curvature flows is set in the context of integral currents for several reasons. First, when sets are regarded as currents, the identifications we want are automatically made. Second, we are able to utilize constructions and estimates proved for currents rather than attempting to reprove them without the current structure. Some basic references for currents are Federer's treatise [F], the text [S] by Simon, and the introductory book [M1] by Morgan.

3.1.1. General currents. A general k (dimensional) *current* T in \mathbb{R}^n is a real-valued continuous linear functional on the real vector space of infinitely differentiable differential k forms in \mathbb{R}^n with compact support. If $k \geq 1$, then the boundary of T is the $k - 1$ current ∂T given by setting $\partial T(\omega) = T(d\omega)$ for differential $k - 1$ forms ω ; i.e., Stokes' theorem becomes a definition.

Associated with any infinitely differentiable proper mapping $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ and any k current T in \mathbb{R}^n is the k current $f_* T$ in \mathbb{R}^m defined by setting $f_* T(\omega) = T(f^* \omega)$ for differential k forms ω in \mathbb{R}^m . In particular, $\partial f_* T = f_* \partial T$.

3.1.2. Rectifiable sets and Lipschitz maps. A subset S of \mathbb{R}^n is called k rectifiable, provided that, for each $\varepsilon > 0$, there is a compact continuously differentiable k -dimensional submanifold with boundary M_ε for which $\mathcal{H}^k(S \Delta M_\varepsilon) < \varepsilon$. (Such sets are called (\mathcal{H}^k, k) rectifiable and \mathcal{H}^k measurable in [FF, § 3.2.14].) When R, S are k rectifiable subsets of \mathbb{R}^n and $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is Lipschitz, then the following is true [F, §§ 3.2.19, 3.2.20, 3.2.23, 3.2.29]:

(i) $R \cup S$ and $R \cap S$ are k rectifiable subsets of \mathbb{R}^n and $f(S)$ is a k rectifiable subset of \mathbb{R}^m in case $m \geq k$;

(ii) \mathcal{H}^k almost every p in S admits an (\mathcal{H}^k, k) approximate tangent plane $\tau_S(p) = \text{Tan}^k(\mathcal{H}^k \llcorner S, p)$, which is a k -dimensional linear subspace of \mathbb{R}^n , and f admits an (\mathcal{H}^k, k) approximate differential $\text{ap } Df(p): \tau_S(p) \rightarrow \mathbb{R}^m$ and an (\mathcal{H}^k, k) approximate Jacobian $\text{ap } J_k f(p) = \|\wedge_k \text{ap } Df(p)\|$;

(iii) If $m \geq k$ and $g: \mathbb{R}^m \rightarrow \mathbb{R}$, then

$$\int_S (g \circ f) \text{ap } J_k f d\mathcal{H}^k = \int_{z \in \mathbb{R}^m} g(z) N(f|S, z) d\mathcal{H}^k z;$$

here $N(f|S, z) = \text{Card}(S \cap f^{-1}\{z\})$ is the multiplicity with which $f|S$ assumes the value z ;

(iv) If $m < k$ and $g: S \rightarrow \mathbb{R}$ is $\mathcal{H}^{n-1} \llcorner S$ integrable, then $S \cap f^{-1}\{y\}$ is $k-m$ rectifiable for \mathcal{L}^m almost every $y \in \mathbb{R}^m$, and

$$\int_S g \cdot \text{ap } J_m f d\mathcal{H}^k = \int_{y \in \mathbb{R}^m} \int_{S \cap f^{-1}\{y\}} g d\mathcal{H}^{k-m} d\mathcal{L}^m y$$

(one form of Federer's coarea formula).

Suppose that S is an $n-1$ rectifiable subset of \mathbb{R}^n and that our surface energy integrand Φ is even, i.e. $\Phi(v) = \Phi(-v)$, for each v in accordance with § 2.1.3. In that case, we extend our notion of surface energy by setting

$$\Phi(S) = \int_{x \in S} \Phi(n_S(x)) d\mathcal{H}^{n-1} x,$$

where n_S is any $\mathcal{H}^{n-1} \llcorner S$ measurable unit vector-valued function for which $n_S(x)$ is perpendicular to $\tau_S(x)$ at almost every x .

3.1.3. Rectifiable and integral currents. (See [F, § 4.1.24].) A k current T in \mathbb{R}^n is called *rectifiable*, provided that it can be expressed as

$$T = \mathbf{t}(S, \theta, \sigma)$$

for some bounded k rectifiable set S , some positive integer-valued $\mathcal{H}^k \llcorner S$ summable density function $\theta: S \rightarrow \mathbb{Z}^+$, and some simple unit k vector-valued orientation function $\sigma: S \rightarrow \wedge_k \mathbb{R}^n$ that is compatible with the approximate tangent plane structure of S ; i.e., $\tau_S(p)$ is the linear subspace associated with $\sigma(p)$ for \mathcal{H}^k almost every p in S [F, § 4.1.28]. If ω is a smooth differential k form with compact support, then

$$T(\omega) = \mathbf{t}(S, \theta, \sigma)(\omega) =: \int_{x \in S} \langle \sigma(x), \omega(x) \rangle \theta(x) d\mathcal{H}^k x;$$

this definition also makes sense for unbounded S 's, provided that θ is locally summable. The *variation measure* $\|T\|$ associated with T is given by setting

$$\|T\| = \mathcal{H}^k \wedge \theta, \quad \text{i.e., } \|T\|(A) = \int_{S \cap A}^* \theta d\mathcal{H}^k$$

for each $A \subset \mathbb{R}^n$; here \int^* denotes the upper integral.

The *mass* of T [F, § 4.1.27] is the number

$$\mathbf{M}(T) = \|T\|\mathbb{R}^n = \int_S \theta d\mathcal{H}^k = \sup \{T(\omega) : \|\omega\| \leq 1\} = \sup \{T(\omega) : |\omega| \leq 1\},$$

while the *size* of T is the number

$$\mathbf{S}(T) = \int_S 1 d\mathcal{H}^k = \mathcal{H}^k(S).$$

We say that T is an *integral k current*, provided that T is a rectifiable k current, and, additionally in case $k \geq 1$, ∂T is a rectifiable $k-1$ current. (The term “integral” here refers to the fact that the density function θ takes positive integer values; real rectifiable currents have real-valued density functions.)

Associated in a natural way with any rectifiable k current $T = \mathbf{t}(S, \theta, \sigma)$ in \mathbb{R}^n and any Lipschitz mapping $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the rectifiable k current $f_{\#}T$ in \mathbb{R}^m defined by setting

$$f_{\#}T(\omega) = \lim_{j \rightarrow \infty} T(g_j^* \omega)$$

for any sequence $\{g_j\}_j$ of infinitely differentiable mappings with uniformly bounded Lipschitz constants that converge uniformly to f [F, § 4.1.14]. Furthermore, the following hold:

(i) If

$$\xi(p) = \sum_{x \in S \cap f^{-1}\{p\} \cap \{z: \text{ap } J_k f(x) > 0\}} \theta(x) \frac{\langle \sigma(x), \wedge_k \text{ap } Df(x) \rangle}{|\langle \sigma(x), \wedge_k \text{ap } Df(x) \rangle|}$$

and $R = \mathbb{R}^m \cap \{p: \xi(p) \neq 0\}$ then [F, § 4.1.30]

$$f_{\#}T = \mathbf{t}\left(R, |\xi|, \frac{\xi}{|\xi|}\right).$$

In particular, $R \subset_{n-1} f(S)$ and, if $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is another Lipschitz mapping that agrees with f on S , then $g_{\#}T = f_{\#}T$;

(ii) If $k \geq 1$ and T is an integral k current, then $f_{\#}T$ is an integral current and $\partial f_{\#}T = f_{\#}\partial T$.

If $T = \mathbf{t}(S, \theta, \sigma)$ is a rectifiable k current it is often the case that the support $\text{spt } T$ of T will be considerably larger than S . It follows from [F, §§ 2.10.6, 2.10.17(4)] that one condition that is sufficient so that $\mathcal{H}^k(\text{spt } T \triangle S) = 0$ is that there exist continuous positive functions $a, b: \mathbb{R}^n \rightarrow \mathbb{R}^+$ such that the *density ratios*

$$\Theta^k(\|T\|, p, r) =: \frac{\|T\|\mathbb{B}^n(p, r)}{\alpha(k)r^k}$$

are not less than $a(p)$ whenever $p \in \text{spt } T$ and $0 < r < b(p)$. If such a and b exist, then it follows that

$$\Theta^{*k}(\|T\|, p) =: \limsup_{r \downarrow 0} \frac{\|T\|\mathbb{B}^n(p, r)}{\alpha(k)r^k} \geq a(p)$$

whenever $p \in \text{spt } (T)$; $\Theta^{*k}(\|T\|, p)$ is called the *upper k density of $\|T\|$ at p* .

We can extend the definitions above in obvious ways to define currents that are locally rectifiable or locally integral.

3.1.4. Integral currents of dimension n in \mathbb{R}^n . (See [FF, § 4.5].) The n -dimensional Euclidean current

$$\mathbb{E}^n = \mathbf{t}(\mathbb{R}^n, 1, \mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n)$$

is a locally integral current with $\partial \mathbb{E}^n = 0$, which, when restricted to a bounded \mathcal{L}^n measurable subset K of \mathbb{R}^n , gives the rectifiable n current

$$[K] = \mathbb{E}^n \llcorner K = \mathbf{t}(K, 1, \mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n).$$

If $[K]$ is an integral current, then $[K] \in \mathcal{H}$, and we can write

$$\partial[K] = \mathcal{H}^{n-1} \wedge * \mathbf{n}_K = \mathbf{t}(\partial K, 1, *\mathbf{n}_K)$$

[F, § 4.5.6], as indicated in § 2.1.2. General integral n currents T in \mathbb{R}^n can be represented uniquely as

$$T = \mathbf{t}(A, \theta, \sigma) = \sum_{i=1}^{\infty} [N_i] - \sum_{j=1}^{\infty} [L_j],$$

where

$$N_i = \{x: \sigma(x) = \mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n \text{ and } \theta(x) \geq i\},$$

$$L_j = \{x: \sigma(x) = -\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n \text{ and } \theta(x) \geq j\}.$$

In particular, $N_1 \cap L_1 = \emptyset$, and $N_1 \supset N_2 \supset N_3 \dots$, and $L_1 \supset L_2 \supset L_3 \dots$. Furthermore [F, § 4.5.17],

$$\partial T = \sum_i \partial[N_i] - \sum_j \partial[L_j]$$

with

$$\|\partial T\| = \sum_i \|\partial[N_i]\| + \sum_j \|\partial[L_j]\|$$

so that

$$\mathbf{M}(\partial T) = \sum_i \mathbf{M}(\partial[N_i]) + \sum_j \mathbf{M}(\partial[L_j]).$$

For such T , we set in the obvious way

$$\Phi(\partial T) = \sum_i \Phi(\partial[N_i]) + \sum_j \Phi(-\partial[L_j]).$$

We note that whenever $[K], [L] \in \mathcal{H}$ then

$$[K] + [L] = [K \cup L] + [K \cap L], \quad \partial[K] + \partial[L] = \partial[K \cup L] + \partial[K \cap L],$$

and we confirm that

$$\partial(K \cup L) \cup \partial(K \cap L) \subset_{n-1} \partial K \cup \partial L$$

and

$$\Phi(\partial[K \cup L]) + \Phi(\partial[K \cap L]) \leq \Phi(\partial[K]) + \Phi(\partial[L]).$$

3.1.5. A compactness theorem for integral currents. (See [F, § 4.2.17], [A2, § 4.3].) There are various compactness theorems for integral and real rectifiable currents. One that is suitable for our purposes is the following. Suppose that T_1, T_2, T_3, \dots are

integral k currents in \mathbb{R}^n such that $\sup_i \mathbf{M}(T_i) < \infty$, $\sup_i \mathbf{M}(\partial T_i) < \infty$, and $\bigcup_i \text{spt}(T_i)$ is bounded. Then there is a subsequence $i(1), i(2), i(3), \dots$ of $1, 2, 3, \dots$ and an integral k current T such that $T_{i(j)} \rightarrow T$ weakly as $j \rightarrow \infty$, i.e., $T_{i(j)}(\omega) \rightarrow T(\omega)$ as $j \rightarrow \infty$ for each smooth differential k from ω . It follows that $\mathbf{M}(T) \leq \liminf_{j \rightarrow \infty} \mathbf{M}(T_{i(j)})$ and is also true that $\mathbf{S}(T) \leq \liminf_{j \rightarrow \infty} \mathbf{S}(T_{i(j)})$ [A2, § 2.10]. In case $k = n$, then $\mathbf{M}(T - T_{i(j)}) \rightarrow 0$; a consequence of this is that, if $T_i = [K_i] \in \mathcal{H}$ for each i , then $T = [K] \in \mathcal{H}$ for some K . In case $k = n - 1$, then $\mathbb{G}(T - T_{i(j)}) \rightarrow 0$ and $\Phi(T) \leq \liminf_{j \rightarrow \infty} \Phi(T_{i(j)})$. (A corresponding lower semicontinuity estimate holds for integrals of appropriate convex surface energy integrands for general dimensions k .)

3.1.6. The cone construction. (See [F, § 4.1.11].) Associated with any integral k , current T ($k \leq n - 1$), and point p in \mathbb{R}^n is the cone $[p] \times T$ over T with vertex p , which is an integral $k + 1$ current with

$$\partial([p] \times T) = T - [p] \times \partial T.$$

In case $\text{spt } T \subset \mathbb{B}^n(p, r)$, then $\text{spt}([p] \times T) \subset \mathbb{B}^n(p, r)$, and we can estimate

$$\mathbf{S}([p] \times T) \leq \mathbf{M}([p] \times T) \leq \frac{r}{k+1} \mathbf{M}(T).$$

As a consequence of the constancy theorem [F, § 4.1.7], together with the cone construction, we infer that, corresponding to each integral $n - 1$ current R with $\partial R = 0$, there is a unique n current $Q = [0] \times R$ with $\partial Q = R$.

3.1.7. Isoperimetric inequalities. (See [A3].) Whenever T is an integral k current (with $1 \leq k \leq n - 1$) with $\partial T = 0$, there is an integral $k + 1$ current Q with $\partial Q = T$, having support in the convex hull of T such that

$$\mathbf{M}(Q) \leq \gamma(k+1) \mathbf{M}(T) \mathbf{S}(T)^{1/k} \leq \gamma(k+1) \mathbf{M}(T)^{(k+1)/k},$$

here the isoperimetric constant $\gamma(k+1)$ is determined by the requirement that equality results if Q is a flat $k+1$ disk having a standard round k sphere T as boundary. As noted in § 3.1.5, Q will be uniquely determined by T in case $k = n - 1$. The left-hand inequality above also holds for real rectifiable currents.

3.1.8. Slicing integral currents by Lipschitz functions. (See [F, § 4.3].) There is a quite general theory of slicing for currents. The special case we need is the following. Suppose that $T = \mathbf{t}(S, \theta, \sigma)$ is an integral k current ($k \geq 1$) and $\rho: \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies a Lipschitz condition with constant 1. We set $m(r) = \|T\|(\{x: \rho(x) < r\})$ for each r so that m is nondecreasing and hence differentiable for almost every r . Then, for almost every r , the following will hold:

- (i) $\|T\|(\{x: \rho(x) = r\}) = 0$;
- (ii) The restriction current $T \llcorner \{x: \rho(x) < r\}$ is an integral k current, while the restriction current $\partial T \llcorner \{x: \rho(x) < r\}$ is an integral $k - 1$ current;
- (iii) The slice current $\langle T, \rho, r \rangle = \mathbf{t}(S \cap \{x: \rho(x) = r\}, \theta, \tau)$, for the appropriate τ , is a well-defined integral $k - 1$ current with $\mathbf{M}(\langle T, \rho, r \rangle) \leq m'(r)$ and

$$\partial(T \llcorner \{x: \rho(x) < r\}) = \partial T \llcorner \{x: \rho(x) < r\} + \langle T, \rho, r \rangle;$$

- (iv) The slice current $\langle \partial T, \rho, r \rangle$ is a well-defined integral $k - 2$ current with

$$\langle \partial T, \rho, r \rangle = -\langle \partial T, \rho, r \rangle.$$

3.1.9. Half-space comparisons. Suppose that $[K] \in \mathcal{H}$ and $\lambda: \mathbb{R}^n \rightarrow \mathbb{R}$ is linear with $\|\lambda\| = 1$. Then, for every r in \mathbb{R} , $[K \cap \{x: \lambda(x) < r\}] \in \mathcal{H}$, and we can use Stokes' theorem and Jensen's inequality [F, § 2.4.19] to infer that

$$\mathbf{M}(\partial[K \cap \{x: \lambda(x) < r\}]) \leq \mathbf{M}(\partial[K])$$

and

$$\Phi(\partial[K \cap \{x: \lambda(x) < r\}]) \leq \Phi(\partial[K]).$$

3.2. Existence of \mathcal{E} minimizers. Suppose that $\Delta t > 0$, $[K_0] \in \mathcal{H}$ and $[L_1], [L_2], [L_3], \dots$ are currents in \mathcal{H} for which

$$\lim_{i \rightarrow \infty} \mathcal{E}([K_0], [L_i], \Delta t) = \inf \{ \mathcal{E}([K_0], [L], \Delta t) : [L] \in \mathcal{H} \}.$$

In view of § 3.1.9, we can modify our L_i 's if necessary so that $\cup_i L_i$ is bounded. Since $\Phi_0 > 0$ (§ 2.1.3), we infer that $\sup_i \mathbf{M}(\partial L_i) < \infty$. The compactness theorem indicated in § 3.1.5 and the lower semicontinuity of Φ integrals guarantee the existence of a subsequence $i(1), i(2), i(3), \dots$ of $1, 2, 3, \dots$ and $[K] \in \mathcal{H}$ such that

$$\lim_{j \rightarrow \infty} \mathbb{G}(\partial[L_{i(j)}] - \partial[K]) = \lim_{j \rightarrow \infty} \mathcal{L}^n(L_{i(j)} \triangle K) = 0$$

and

$$\mathcal{E}([K_0], [K], \Delta t) = \inf \{ \mathcal{E}([K_0], [L], \Delta t) : [L] \in \mathcal{H} \}.$$

We define such a $[K]$ to be an \mathcal{E} minimizer for $[K_0]$ over Δt . (If we say $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt , it is to be presumed that $[K_0]$ and $[K]$ are in \mathcal{H} and $\Delta t > 0$.) We infer from § 3.1.5 again that the support of any such \mathcal{E} minimizer $[K]$ must lie within the convex hull of the closure of ∂K_0 .

3.3. Properties of \mathcal{E} minimizers. Suppose that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt .

3.3.1. For $[L] \in \mathcal{H}$, we check

$$K_0 \triangle L \sim K_0 \triangle K \subset K \triangle L$$

and infer from the definition of \mathcal{E} that

$$\Phi(\partial[K]) \leq \Phi(\partial[L]) + \frac{1}{\Delta t} \int_{x \in K \triangle L} \text{dist}(x, \partial K_0) d\mathcal{L}^n x.$$

3.3.2. We assert that, whenever Q is an integral n current in \mathbb{R}^n , then

$$\Phi(\partial[K]) \leq \Phi(\partial[K] + \partial Q) + \mathbf{S}(Q) \sup \left\{ \frac{\text{dist}(x, \partial K_0)}{\Delta t} : x \in \text{spt } Q \right\}.$$

To see this, we write

$$Q = \sum_i [Q_i] - \sum_j [P_j], \quad [K] + Q = \sum_i [N_i] - \sum_j [J_j],$$

as in § 3.1.4. As a comparison region, we set $L = N_1 = (K \sim P_1) \cup Q_1$ so that

$$K \triangle L = (Q_1 \sim K) \cup (P_1 \cap K) \subset Q_1 \cup P_1.$$

Our assertion now follows from §§ 3.3.1 and 3.1.3.

3.3.3. Suppose that $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is Lipschitz and $W = \{x: \varphi(x) \neq x\}$ is bounded. We assert the following:

$$(i) \quad \Phi_0 \mathcal{H}^{n-1}(\partial K \cap W) \leq \Phi^0 \mathcal{H}^{n-1}(\varphi(\partial K \cap W)) + E,$$

and, in case Φ is even,

$$(ii) \quad \Phi(\partial K \cap W) \leq \Phi(\varphi(\partial K \cap W)) + E,$$

where

$$E = \frac{\text{diam}(W \cup \varphi(W))}{n} [\mathcal{H}^{n-1}(\partial K \cap W) + \mathcal{H}^{n-1}(\varphi(\partial K \cap W))] \sup_{x \in \text{spt } Q} \left\{ \frac{\text{dist}(x, \partial K_0)}{\Delta t} \right\}.$$

To see this, we write

$$\varphi_*[K] - [K] = \sum_i [N_i] - \sum_j [L_j]$$

as in § 3.1.4, and, as a comparison region, we take $J = N_1 \cup (K \sim L_1)$ so that (since $N_1 \cap L_1 = \emptyset$)

$$K \triangle J = (N_1 \sim K) \cup (K \cap L_1) \subset N_1 \cup L_1$$

with

$$\partial(N_1 \cup L_1) \subset_{n-1} \partial N_1 \cup \partial L_1 \subset_{n-1} \varphi(\partial K \cap W) \cup (\partial K \cap W).$$

We then infer from the cone construction with p in W that

$$\begin{aligned} \mathcal{L}^n(K \triangle J) &\leq \mathcal{L}^n(N_1 \cup L_1) \\ &= \mathbf{M}([p] \times \partial[N_1 \cup L_1]) \\ &\leq \frac{\text{diam}(W \cup \varphi W)}{n} [\mathcal{H}^{n-1}(\partial K \cap W) + \mathcal{H}^{n-1}(\varphi(\partial K \cap W))]. \end{aligned}$$

We write

$$\varphi_*[K] = [K] + \sum_i [N_i] - \sum_j [L_j] = [J] + \sum_{i \geq 2} [N_i^*] - \sum_j [L_j^*]$$

for appropriate N_i^* 's and L_j^* 's, and we use § 3.1.4 to conclude that

$$\partial J \subset_{n-1} \varphi(\partial K \cap W) \cup (\partial K \sim W).$$

Since $\varphi(x) = x$ for $x \in \partial K \sim W$, we infer from the formula in § 3.1.3(i) that $\mathbf{n}_J(p) = \mathbf{n}_K(p)$ for \mathcal{H}^{n-1} almost every point p in $\partial K \cap \partial J \sim \varphi(\partial K \cap W)$. *Caution:* Examples show that there can be a set of positive \mathcal{H}^{n-1} measure in $\partial K \cap \partial J \cap \varphi(\partial K \cap W)$ for which $\mathbf{n}_J(p) = -\mathbf{n}_K(p)$. We infer that

$$\begin{aligned} \Phi(\partial J) &= \Phi(\partial J \llcorner (\partial K \sim W)) + \Phi(\partial J \llcorner \varphi(\partial K \cap W)) \\ &= \Phi(\partial K \llcorner (\partial K \sim W)) + \Phi(\partial J \llcorner \varphi(\partial K \cap W)). \end{aligned}$$

In view of § 3.3.1, we estimate that

$$\begin{aligned} \Phi(\partial K \llcorner W) &= \Phi(\partial K) - \Phi(\partial K \llcorner (\partial K \sim W)) \\ &\leq \Phi(\partial J \llcorner \varphi(\partial K \cap W)) + \frac{1}{\Delta t} \int_{x \in K \triangle J} \text{dist}(x, \partial K_0) d\mathcal{L}^n x. \end{aligned}$$

The two assertions of § 3.3.3 follow.

3.4. \mathcal{E} minimizers have lower density ratio bounds. Suppose that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt . Suppose also that the convex hull of K_0 has diameter no larger than $D-1$ (for some $D \geq 1$). We will show the existence of uniform lower bounds to the density ratios of the measures $\|\partial[K]\|$ at points p in \mathbb{R}^n for which $\Theta^{*n-1}(\|\partial[K]\|, p) > 0$. As noted in § 3.1.3, this will imply that $\mathcal{H}^{n-1}(\text{closure}(\partial K) \triangle \partial K) = 0$.

We abbreviate $T = \partial[K]$ and suppose that $p \in \mathbb{R}^n$ with $\Theta^{*n-1}(\|T\|, p) > 0$. We set $\rho(x) = |x - p|$ for $x \in \mathbb{R}^n$ and $T_r = T \llcorner \{x: \rho(x) < r\}$,

$$m(r) = \mathbf{M}(T_r) = \mathcal{H}^{n-1}(\partial K \cap \mathbb{B}^n(p, r))$$

for $r > 0$. We use § 3.1.8 to infer that for almost every $r > 0$, the slice $\langle T, \rho, r \rangle$ exists with $\partial T_r = \langle T, \rho, r \rangle$ and $\mathbf{M}(\partial T_r) \leq m'(r)$. We then use the isoperimetric inequality in § 3.1.7 to infer the existence of an integral $n-1$ current R having support in $\mathbb{B}^n(p, r)$ such that $\partial R = \partial T_r = \langle T, \rho, r \rangle$ and

$$\mathbf{M}(R) \leq \gamma(n-1)\mathbf{M}(\partial T_r)^{(n-1)/(n-2)} \leq \gamma(n-1)m'(r)^{(n-1)/(n-2)}.$$

Since $\partial(R - T_r) = 0$, we infer from § 3.1.6 that the cone $Q = [p] \rtimes (R - T_r)$ satisfies the conditions $\partial Q = R - T_r$ and

$$\mathbf{M}(Q) \leq \frac{r}{n} \mathbf{M}(R - T_r) \leq \frac{r}{n} (\gamma(n-1)m'(r)^{(n-1)/(n-2)} + m(r)).$$

We infer that

$$\Phi(T + \partial Q) - \Phi(T) = \Phi(R + (T - T_r)) - \Phi(T_r + (T - T_r)) = \Phi(R) - \Phi(T_r),$$

and we use § 3.3.2 (recalling that $T = \partial[K]$) to conclude that

$$\begin{aligned} \Phi_0 m(r) &\leq \Phi(T_r) \\ &\leq \Phi(R) + \mathbf{M}(Q) \sup \left\{ \frac{\text{dist}(x, \partial K_0)}{\Delta t} : x \in \mathbb{B}^n(p, r) \right\} \\ &\leq \Phi^0 \gamma(n-1)m'(r)^{(n-1)/(n-2)} \\ &\quad + \frac{r}{n} (\gamma(n-1)m'(r)^{(n-1)/(n-2)} + m(r)) \sup \left\{ \frac{\text{dist}(x, \partial K_0)}{\Delta t} : x \in \mathbb{B}^n(p, r) \right\}. \end{aligned}$$

We restrict r 's to not exceed 1, and we infer that

$$m(r) \leq \frac{\Phi^0}{\Phi_0} \gamma(n-1)m'(r)^{(n-1)/(n-2)} + r \frac{D}{n\Delta t \Phi_0} (\gamma(n-1)m'(r)^{(n-1)/(n-2)} + m(r)),$$

so that

$$m(r) \left(1 - r \left(\frac{D}{n\Delta t \Phi_0} \right) \right) \leq \left(\gamma(n-1) \frac{\Phi^0}{\Phi_0} m'(r)^{(n-1)/(n-2)} \right) \left(1 + r \left(\frac{D}{n\Delta t \Phi_0} \right) \right).$$

In case

$$r \leq r_0 = \left(\frac{n\Phi_0}{3D} \right) \Delta t \quad \text{and} \quad C^{(n-1)/(n-2)} = 2\gamma(n-1) \frac{\Phi^0}{\Phi_0},$$

then

$$m(r) \leq C^{(n-1)/(n-2)} m'(r)^{(n-1)/(n-2)},$$

which implies that

$$((n-1)m(r)^{1/(n-1)})' = \frac{m'(r)}{m(r)^{(n-2)/(n-1)}} \geq \frac{1}{C};$$

since $m(r)$ is nondecreasing in r , we conclude that

$$(n-1)m(r)^{1/(n-1)} \geq \frac{r}{C}, \quad m(r) \geq \left(\frac{1}{(n-1)C} \right)^{n-1} r^{n-1}.$$

If $r_0 \leq r \leq \Delta t$, we can conclude further that

$$\frac{m(r)}{r^{n-1}} \geq \frac{m(r_0)}{r_0^{n-1}} \left(\frac{r_0}{r} \right)^{n-1} \geq \left(\frac{1}{(n-1)C} \right)^{n-1} \left(\frac{n\Phi_0}{3D} \right)^{n-1}.$$

We set

$$\theta = \frac{1}{(n-1)^{n-1}} \left(\frac{\Phi_0}{2\gamma(n-1)\Phi^0} \right)^{(n-1)^2/(n-2)} \inf \left\{ 1, \frac{n\Phi_0}{3D} \right\},$$

and we conclude that

$$\frac{m(r)}{r^{n-1}} \geq \theta \quad \text{for } 0 < r \leq \Delta t.$$

3.5. \mathcal{E} minimizers are (Φ, ω, δ) minimal currents in the sense of Bombieri. (See [Bo, Def. 1].) Suppose that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt and $D-1 \geq \text{diam}(K_0)$ (for some $D > 1$). Suppose that X is any $n-1$ integral current with $\partial X = 0$ whose support lies in a compact set $C \in \mathcal{C}$ for which

$$\text{diam}(C) = r \leq \delta =: \frac{n\Phi_0}{3D} \Delta t.$$

We assert that

$$\Phi(\partial[K] \llcorner C) \leq \Phi(\partial[K] \llcorner C + X) + \omega(r) \mathbf{M}(\partial[K] \llcorner C + X),$$

where

$$\omega(r) = r \left(\frac{3D\Phi^0}{n\Phi_0\Delta t} \right).$$

This condition is what is required for the current $\partial[K]$ to be (Φ, ω, δ) *minimal* in the sense of Definition 1 of [Bo].

In case Φ is smooth and elliptic, we assert further the existence of an open subset U of \mathbb{R}^n such that $\mathcal{H}^{n-1}(\text{spt } \partial[K] \sim U) = 0$ and $\text{spt } \partial[K] \cap U$ is a continuously differentiable $(n-1)$ -dimensional submanifold of \mathbb{R}^n . In case $n=3$, then $\text{spt } \partial[K]$ is a compact continuously differentiable two-dimensional submanifold of \mathbb{R}^3 without boundary.

To see our first assertion, we pick p in C and set $Q = [p] \times X$ (see § 3.1.6) so that $\partial Q = X$ and

$$\mathbf{S}(Q) \leq \mathbf{M}(Q) \leq \frac{r}{n} \mathbf{M}(X) \leq \frac{r}{n} [\mathbf{M}(\partial[K] \llcorner C + X) + \mathbf{M}(\partial[K] \llcorner C)].$$

From § 3.3.2, we infer that $\Phi(\partial[K]) \leq \Phi(\partial[K] + X) + \mathbf{S}(Q)(D/\Delta t)$, so that

$$\Phi(\partial[K] \llcorner C) \leq \Phi(\partial[K] \llcorner C + X) + \frac{rD}{n\Delta t\Phi_0} [\Phi(\partial[K] \llcorner C + X) + \Phi(\partial[K] \llcorner C)].$$

In case $r \leq \delta$, we infer that

$$\frac{1 + \frac{rD}{n\Delta t\Phi_0}}{1 - \frac{rD}{n\Delta t\Phi_0}} \leq 1 + 3 \frac{rD}{n\Delta t\Phi_0},$$

and hence

$$\Phi(\partial[K] \llcorner C) \leq \Phi(\partial[K] \llcorner C + X) \left(1 + 3 \frac{rD}{n\Delta t\Phi_0} \right).$$

Our first assertion follows directly.

The almost everywhere regularity asserted above follows from the final remark in [Bo], together with the density ratio estimate of § 3.4. The everywhere regularity in the case where $n = 3$ similarly follows, with the additional observation that tangent cones to $\partial[K]$ are absolutely Φ minimizing, hence everywhere regular in accordance with [ASS].

3.6. The supports of \mathcal{E} minimizers are (γ, δ) restricted sets in the sense of Almgren. (See [A1, § II.1].) Suppose $\Delta t > 0$ and $[K_0], [K]$ are solids in \mathcal{K} with

$$\mathcal{E}([K_0], [K], \Delta t) = \inf \{ \mathcal{E}([K_0], [L], \Delta t) : [L] \in \mathcal{K} \}$$

and $D - 1 \geq \text{diam}(K_0)$ (for some $D > 1$). We set

$$S = \text{spt } \partial[K], \quad \gamma = 2 \frac{\Phi^0}{\Phi_0}, \quad \delta = \inf \left\{ 1, \frac{n\Phi_0}{D} \Delta t \right\},$$

and assert

- (i) $\mathcal{H}^{n-1}(S \triangle \partial K) = 0$;
- (ii) $\mathcal{H}^{n-1}(S \cap W) \leq \gamma \mathcal{H}^{n-1}(\varphi(S \cap W))$ whenever
 - (a) $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a Lipschitz mapping,
 - (b) $W = \{x : \varphi(x) \neq x\}$,
 - (c) $\text{diam}(W \cup \varphi(W)) < \delta$.

These conditions comprise the definition of S being (γ, δ) *restricted* with respect to the empty set in the sense of [A1, § II.1]. This implies that S has the various properties demonstrated in Theorem II.3 of [A1].

Assertion (i) follows from the lower density ratio bounds established in § 3.4, as noted in § 3.3.1. Assertion (ii) follows from assertion (i) and § 3.3.3(i); compare the estimates in § 3.4 and § 3.7, below.

3.7. The supports of \mathcal{E} minimizers are $(\Phi, \varepsilon, \delta)$ minimal sets in the sense of Almgren. (See [A1, § III.1].) Suppose that Φ is an even integrand, $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt , $D \geq 1 + \text{diam}(K_0)$, and $\lambda = 3D/n\Phi_0$. We set

$$S = \text{spt } \partial[K], \quad \varepsilon(r) = r \left(\frac{\lambda}{\Delta t} \right), \quad \delta = \inf \left\{ 1, \frac{\Delta t}{\lambda} \right\}$$

for $0 < r < \delta$ and assert

- (i) S is (γ, δ) restricted (for suitable $\gamma > 1$);
- (ii) $\Phi(S \cap W) \leq (1 + \varepsilon(r))\Phi(\varphi(S \cap W))$ whenever
 - (a) $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a Lipschitz mapping,
 - (b) $W = \{x : \varphi(x) \neq x\}$,
 - (c) $\text{diam}(W \cup \varphi(W)) = r < \delta$.

These conditions comprise the definition of S being $(\Phi, \varepsilon, \delta)$ *minimal* with respect to the empty set in the sense of [A1, § III.1]. This implies that S has the various properties demonstrated in Theorem III.3 of [A1]; in particular, if Φ is smooth and elliptic, then there is an open subset U of \mathbb{R}^n such that $\mathcal{H}^{n-1}(S \sim U) = 0$ and $S \cap U$ is locally a Hölder continuously differentiable $(n-1)$ -dimensional submanifold of \mathbb{R}^n for each Hölder exponent less than $\frac{1}{2}$.

Assertion (i) follows from § 3.6. Assertion (ii) follows from § 3.3.3(ii). In particular, we estimate from assertion (i), above, and § 3.3.3(ii) that

$$\Phi(S \cap W) \leq \Phi(\varphi(S \cap W)) + \frac{r}{n\Phi_0} [\Phi(S \cap W) + \Phi(\varphi(S \cap W))] \frac{D}{\Delta t},$$

so that

$$\Phi(S \cap W) \left(1 - r \frac{\lambda}{3\Delta t}\right) \leq \Phi(\varphi(S \cap W)) \left(1 + r \frac{\lambda}{3\Delta t}\right).$$

We check that

$$\frac{\left(1 + r \frac{\lambda}{3\Delta t}\right)}{\left(1 - r \frac{\lambda}{3\Delta t}\right)} \leq \left(1 + r \frac{\lambda}{\Delta t}\right)$$

provided $r \leq \Delta t / \lambda$.

THEOREM 3.8 (higher regularity of \mathcal{E} minimizers). *Suppose that Φ is smooth and elliptic and that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt . For each $p \in \mathbb{R}^n$, let $g(p) = \pm \text{dist}(p, \partial K_0)$, where we take the negative sign if $p \in K_0$, and take the positive sign otherwise.*

(1) *For \mathcal{H}^{n-1} almost every point p in $\text{spt } \partial[K]$, there is a Euclidean motion $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$, together with positive numbers R and M and a continuously differentiable function $f: \mathbb{B}^{n-1}(0, 2R) \rightarrow (-M, M)$ for which*

$$L(K) \cap (\mathbb{B}^{n-1}(0, 2R) \times (-M, M)) = \{(x, y): x \in \mathbb{B}^{n-1}(0, 2R), -M < y < f(x)\}.$$

Furthermore, g is Lipschitz continuous with Lipschitz constant equal to 1, and $g(x, f(x))$ is Lipschitz as a function of x .

For simplicity of notation, we will hereafter assume that L is the identity mapping on \mathbb{R}^n and that R , M , and f are fixed.

(2) *f weakly satisfies*

$$(3.8.1) \quad \sum_{i=1}^{n-1} \frac{\partial}{\partial x_i} \left(\frac{\partial \Phi}{\partial v_i}(v) \right) \Big|_{v=(-\partial f / \partial x_1)(x), \dots, -(\partial f / \partial x_{n-1})(x), 1)} = \frac{g(x, f(x))}{\Delta t},$$

is thus twice Hölder continuously differentiable in $\mathbb{B}^{n-1}(0, R)$, and strongly satisfies

$$(3.8.2) \quad \begin{aligned} \frac{g(x, f(x))}{\Delta t} &= - \sum_{i,j=1}^{n-1} \frac{\partial^2 \Phi}{\partial v_i \partial v_j}(v) \Big|_{v=(-\partial f / \partial x_1)(x), \dots, -(\partial f / \partial x_{n-1})(x), 1)} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &= H_{\Phi, \partial K}(x, f(x)) \cdot \mathbf{n}_K(x, f(x)). \end{aligned}$$

(3) f , in fact, satisfies the *a priori* estimates on second derivatives

$$\begin{aligned} & \sup \left\{ \left| \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right| : 1 \leq i \leq j \leq n-1, x \in \mathbb{B}^{n-1}(0, R) \right\} \\ & \leq C \left(\frac{1}{R} \sup \left\{ \left| \frac{\partial f}{\partial x_i}(x) \right| : 1 \leq i \leq n-1, x \in \mathbb{B}^{n-1}(0, 2R) \right\} \right. \\ & \quad \left. + \frac{R}{\Delta t} \sup \{ |\nabla g(x)| : x \in \mathbb{B}^{n-1}(0, 2R) \} \right); \end{aligned}$$

here C is a constant depending only on an upper bound for $|\nabla f|$.

(4) \mathcal{H}^{n-1} almost everywhere, $\text{spt } \partial[K]$ is a twice continuously differentiable submanifold of \mathbb{R}^n , with estimates on its principal curvatures analogous to those in (3) above.

(5) If $|g|$ is sufficiently small, g will be as differentiable as ∂K_0 is. If g is k times Hölder continuously differentiable with exponent α for some positive integer k and some $0 < \alpha < 1$, and if f is a Lipschitz solution to (3.8.1), then f is, in fact, $k+2$ times Hölder continuously differentiable with exponent α .

Proof. (1) We know from our analysis in § 3.5 or § 3.7, above, that $\text{spt } \partial[K]$ is almost everywhere a continuously differentiable submanifold of \mathbb{R}^n ; the distance function is always Lipschitz with Lipschitz constant 1 (by the triangle inequality), and the Lipschitz dependence on x follows from the fact that f is Lipschitz.

(2) Since $[K]$ is an \mathcal{E} minimizer, f must weakly satisfy the Euler–Lagrange equation for \mathcal{E} , which is (3.8.1). Since f is Lipschitz, (3.8.1) is uniformly elliptic; i.e.,

$$\sum_{i,j=1}^{n-1} \partial_{v_i}(\partial_{v_j} \Phi) \xi_i \xi_j \geq \nu |\xi|^2$$

for every ξ in \mathbb{R}^{n-1} and some $\nu = \nu(\|f\|_{\text{Lip}}) > 0$.

Using [LU, Chap. 6, Thm. 1.1, p. 339], we conclude that f is Hölder continuously differentiable with Hölder norm depending only on the sup norm of ∇f . However, we can differentiate (3.8.1) again and conclude that the first partial derivative of f with respect to x_k is itself a solution of the linear PDE

$$\frac{1}{\Delta t} \left(\frac{\partial g}{\partial x_k}(x, f(x)) + \frac{\partial g}{\partial y}(x, f(x)) \frac{\partial f}{\partial x_k}(x) \right) = \sum_{i,j=1}^{n-1} \frac{\partial}{\partial x_i} \left(a_{i,j}(x) \frac{\partial (f/\partial x_k)}{\partial x_j}(x) \right)$$

for

$$a_{i,j} = \frac{\partial^2 \Phi}{\partial v_i \partial v_j} \left(-\frac{\partial f}{\partial x_1}, \dots, -\frac{\partial f}{\partial x_{n-1}}, 1 \right),$$

which is Hölder continuous. Using [G, Chap. VII, Thm. 1.2, p. 219] (and the fact that g is Lipschitz), we conclude that $\partial f/\partial x_k$ is also Hölder continuously differentiable. Conclusion (3) now is a consequence of the fact that $\partial f/\partial x_k$ satisfies a linear PDE.

(3) This is essentially a restatement of (1)–(3).

(4) This follows from [KP] and [G] as before. \square

Comment. We do not use the estimates on the second derivatives in (3) in this paper, nor do we use the curvature estimates in (4) or the estimates in (5). These results are included here as essentially the maximum regularity that we know, in general, about \mathcal{E} minimizers (see § 7). We note that it is difficult to use (5) to conclude higher differentiability in the approximate flows $\partial[K_j(t)]$ constructed as in § 2.6, since any given $\Delta t = 2^{-j}$ may not be sufficiently small. Conclusions (1), (3), and (5) of Theorem 5.4 give our best general barrier estimates; see also the barriers constructed in 7.1.

THEOREM 3.9 (scaling of \mathcal{E} minimizers and tangent currents). *Suppose that our surface energy function Φ is smooth and elliptic and the $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt . For each point p in \mathbb{R}^n (which we fix) and each scale factor $R > 0$, we set $f_{p,R}(x) = R(x - p)$ (a translation followed by a homothety) and $[K(R)] = [f_{p,R}K]$ and $[K_0(R)] = [f_{p,R}K_0]$.*

(1) *Each $[K(R)]$ is an \mathcal{E} minimizer for $[K_0(R)]$ over $R^2\Delta t$, i.e.,*

$$\begin{aligned} & \Phi(\partial[K(R)]) + \frac{1}{R^2\Delta t} \int_{x \in K(R) \triangle K_0(R)} \text{dist}(x, \partial K_0(R)) \, d\mathcal{L}^n x \\ &= \inf \left\{ \Phi(\partial[L]) + \frac{1}{R^2\Delta t} \int_{x \in L \triangle K_0(R)} \text{dist}(x, \partial K_0(R)) \, d\mathcal{L}^n x : [L] \in \mathcal{H} \right\}. \end{aligned}$$

(2) *Each increasing sequence $R(1), R(2), R(3), \dots$ of positive numbers that diverges to infinity has a subsequence $R(i(1)), R(i(2)), R(i(3)), \dots$ such that the currents $[K(R(i(j)))]$ converge, as $j \rightarrow \infty$, to a limit locally integral current $[J]$; such convergence means that, for each radius r , $\lim_{j \rightarrow \infty} \mathcal{L}^n(J \triangle K(R(i(j)))) = 0$. The boundary $\partial[J]$ of each such limit current $[J]$ is absolutely Φ minimizing, and the supports of the $\partial[K(R(i(j)))]$'s converge to the support of $\partial[J]$, as $j \rightarrow \infty$, locally uniformly in the Hausdorff distance topology.*

Notation. Any such current $\partial[J]$ is called a *tangent current* to $\partial[K]$ at p . Conclusion (2) asserts that any tangent current to any \mathcal{E} minimizer at any point is locally Φ minimizing. In general, it is an open question whether there can be more than one tangent current at a given point p .

Proof. With regard to the first assertion above, we compute that

$$\begin{aligned} \Phi(\partial[K_0(R)]) &= R^{n-1}\Phi(\partial[K]), \\ \frac{1}{\Delta t} \int_{x \in K(R) \triangle K_0(R)} \text{dist}(x, \partial K_0(R)) \, d\mathcal{L}^n x &= R^{n+1} \frac{1}{\Delta t} \int_{x \in K \triangle K_0} \text{dist}(x, \partial K_0) \, d\mathcal{L}^n x. \end{aligned}$$

The existence of the convergent sequence $\{K(R(i(j)))\}_j$ follows from general compactness theorems for locally integral currents such as [A2, § 4.4] (see also [F, § 4.2.17], [ASS, Thm. 1.1, p. 225], and §§ 3.1.5, 3.2, above). The bounds on mass necessary for applying such theorems follow from the observation that, for each q and r ,

$$\Phi(\partial[K(R)]) \leq \Phi(\partial[K^R \sim \mathbb{B}^n(q, r)]) + \frac{1}{R^2\Delta t} \int_{x \in \mathbb{B}^n(q, r)} \text{dist}(x, \partial K_0(R)) \, d\mathcal{L}^n x,$$

which implies that

$$\Phi(\partial[K(R)] \llcorner \mathbb{B}^n(q, r)) \leq \Phi^0 n \alpha(n) r^{n-1} + \frac{1}{R^2\Delta t} \int_{\mathbb{B}^n(q, r)} \text{dist}(x, \partial K_0(R)) \, d\mathcal{L}^n x,$$

since $\mathcal{H}^{n-1}(\partial \mathbb{B}^n(q, r)) = n \alpha(n) r^{n-1}$. For each fixed q and r , we can check that

$$\lim_{R \rightarrow \infty} \frac{1}{R^2\Delta t} \int_{\mathbb{B}^n(q, r)} \text{dist}(x, \partial K_0(R)) \, d\mathcal{L}^n x = 0.$$

This last fact, in addition to providing upper density ratio bounds, implies that the integrals over $K(R) \triangle K_0(R)$ matter less in the \mathcal{E} minimization as R becomes large. We use this fact to infer that any limit $\partial[J]$ is locally Φ minimizing.

The lower density ratio bounds established in § 3.4 are used to establish the convergence of the supports. \square

THEOREM 3.10 (elliptic \mathcal{E} minimizers are smooth near contact points with smooth barriers and in low dimensions). *Suppose that our surface energy function Φ is smooth and elliptic and that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt .*

(1) *If $\{p\} = \text{spt } \partial[K] \cap \mathbb{B}^n(q, r)$ for some points p and q in \mathbb{R}^n and $r = |q - p|$, then there is an open neighborhood U of p such that $\text{spt } (\partial[K]) \cap U$ is a twice continuously differentiable hypersurface in \mathbb{R}^n . In particular, $\text{spt } \partial[K] \cap U = \partial K \cap U$.*

(2) *If $n = 2$ (respectively, $n = 3$), then $\text{spt } \partial[K] = \partial K$ is everywhere a smooth curve in the plane (respectively, a smooth surface in space).*

Proof. The criterion that we use for regularity above is that the support of $\partial[K]$ within a small ball centered at p can, in fact, be contained in a very thin disk through p in that ball. Knowing this, we can then invoke the regularity construction of Bombieri to conclude C^1 regularity in a smaller ball; the higher regularity then follows from Theorem 3.8(2). The way in which we establish such a disk condition on the supports is by showing that there is a tangent current to $\partial[K]$ at p whose support is regular. This being the case, we also infer that there is a tangent current there having a hyperplane as support. The disk condition is then a consequence of the convergence of the supports given in conclusion (2) of Theorem 3.9.

Now consider the conditions assumed in assertion (1) of Theorem 3.10. We let $f_{p,R}$ be as in Theorem 3.9 (same p) and let $\partial[J]$ be any Φ minimizing tangent current as in Theorem 3.9(2). We infer that the support of $\partial[J]$ lies in the half-space $\{v: v \cdot (q - p) \geq 0\}$ and contains the origin. We now use Remark 1 about Φ minimizers following [ASS, Thm. 1.2, p. 227] to conclude that the origin is a regular point of $\text{spt } \partial[J]$. This then implies conclusion (1), above. The first assertion of conclusion (2) of the above theorem follows in a similar way, based on the one-dimensional regularity for Φ minimizers given in [F, § 5.3.20]. The second assertion of conclusion (2) similarly is based on the two-dimensional regularity of Φ minimizers given in [ASS, Cor. 3.2, p. 255]. \square

4. Existence and Hölder continuity of flat Φ curvature flows. The goal of this section is to establish modulus of continuity estimates on approximate Φ curvature flows that imply a priori Hölder continuity in the \mathbb{G} metric of any limit flat Φ curvature flow. Such a modulus of continuity estimate is a pivotal result in this paper in justifying the variational approach to curvature motion. It follows from our \mathcal{E} minimization construction that each $K_j(t)$ lies within the convex hull of the initial $K(0)$ with

$$\Phi_0 \mathbf{M}(\partial[K_j(t)]) \leq \Phi(\partial[K_j(t)]) \leq \Phi(\partial[K(0)]) \leq \Phi^0 \mathbf{M}(\partial[K(0)]).$$

We can thus use compactness theorems for integral currents (§ 3.1.5), together with Cantor's diagonal process, to infer the existence of a subsequence $j(1), j(2), j(3), \dots$ of $1, 2, 3, \dots$, so that, for a dense set of times, the currents $\partial[K_{j(i)}(t)]$ will converge to a limit current $\partial[K(t)]$ as $i \rightarrow \infty$. Since this fact, in itself, does not preclude the possibility that the $\partial[K(t)]$'s might vary in a wildly discontinuous way, some additional estimate is necessary to guarantee continuity, and the Hölder estimate does this.

Suppose that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt . Since we can always take $K = K_0$ in minimizing $\mathcal{E}([K_0], [K], \Delta t)$, we will take a different K only if there is a reduction in surface energy at least equal to the cost

$$\frac{1}{\Delta t} \int_{p \in K_0 \Delta K} \text{dist}(p, \partial K_0) d\mathcal{L}^n p.$$

In Proposition 4.2 (based on Proposition 4.1), we are effectively able to dominate the volume of $K_0 \triangle K$ in terms of

$$\frac{\Delta t^{1/2}(\Phi(\partial[K_0]) - \Phi(\partial[K]))^{1/2}}{\theta},$$

where θ denotes a positive lower bound to the density ratios of $\|\partial[K_0]\|$; the necessary bounds were established in § 3.4. Theorem 4.4 is a formal statement of the a priori Hölder bound, and Theorem 4.5 is a formal statement of the existence of flat Φ curvature flows.

The lower density ratio bound θ for $\|\partial[K_0]\|$ depends on the assumption that $[K_0]$ itself is an \mathcal{E} minimizer. Without this assumption, $\partial[K_0]$ might contain filigree and nooks and crannies, and hence there may be a large volume within a short distance from ∂K_0 .

We note that the best general estimate we have on how far ∂K can be from ∂K_0 (with no minimizing assumption on K_0) is comparable to $\Delta t^{1/2}$ (which would not give a uniform modulus of continuity when summed over many Δt 's). One side of this bound will be obtained in conclusion (2) of Proposition 5.3; the other side of the bound follows if we reformulate our study in terms of complementary currents as noted in Step 3 of the proof of Theorem 5.4. See also conclusions (1) and (3) of Theorem 5.4.

PROPOSITION 4.1 (a volume-distance inequality). *Suppose that C is a closed subset of \mathbb{R}^n and $\rho(x) = \text{dist}(x, C)$ for each x . Suppose also that A is an \mathcal{L}^n measurable subset of \mathbb{R}^n and that Δt and E are positive numbers such that*

$$\frac{1}{\Delta t} \int_A \rho \, d\mathcal{L}^n \leq E.$$

Then, for each $0 < R < \infty$,

$$\mathcal{L}^n(A \sim C) \leq 2^{1/2} (\sup \{ \mathcal{H}^{n-1}(A \cap \rho^{-1}\{r\}) : 0 < r < R \})^{1/2} (\Delta t)^{1/2} + \frac{\Delta t}{R} E.$$

Proof of Proposition 4.1.

Step 1. For each function $f: [0, R] \rightarrow [0, S]$, we have

$$\int_0^R f \, dr \leq 2^{1/2} S^{1/2} \left(\int_0^R r f(r) \, dr \right)^{1/2}.$$

Proof. If f is not measurable, then neither side of the asserted inequality is defined. Suppose then that f is measurable and let $h: [0, S] \rightarrow [0, R]$ be the distribution function of f defined by requiring that

$$h(s) = \mathcal{L}^1\{r: f(r) > s\}$$

for each s . We denote by $g: [0, R] \rightarrow [0, S]$ the nonincreasing function having h as distribution function; g is called the decreasing rearrangement of f . Fubini's theorem implies that

$$\int_0^R f \, dr = \int_0^S h \, ds = \int_0^R g \, dr,$$

and it is obvious that

$$\int_0^R g \rho \, dr \leq \int_0^R f \rho \, dr$$

for every nondecreasing function $\rho : [0, R] \rightarrow \mathbb{R}^+$. We use this second fact, together with Fubini's theorem, to estimate

$$\begin{aligned} \int_0^R rf(r) dr &\geq \int_0^R rg(r) dr \\ &= \int_{r=0}^R \int_{s=0}^{g(r)} r ds dr \\ &= \int_{s=0}^S \int_{r=0}^{h(s)} r dr ds \\ &= \frac{1}{2} \int_{s=0}^S h^2(s) ds. \end{aligned}$$

We use the first fact, together with the extreme inequality above and Schwarz's inequality, to estimate

$$\begin{aligned} \int_{r=0}^R f(r) dr &= \int_{s=0}^S h(s) ds \\ &\leq S^{1/2} \left(\int_{s=0}^S h^2(s) ds \right)^{1/2} \\ &\leq 2^{1/2} S^{1/2} \left(\int_0^R rf(r) dr \right)^{1/2}, \end{aligned}$$

which establishes the assertion of Step 1.

Step 2. Suppose that $\mathcal{H}^{n-1}(\rho^{-1}\{r\} \cap A) \leq S < \infty$ for \mathcal{L}^1 almost every $0 < r < R$. Then

$$\mathcal{L}^n(A \cap \rho^{-1}(0, R)) \leq 2^{1/2} S^{1/2} \left(\int_{A \cap \rho^{-1}(0, R)} \rho d\mathcal{L}^n \right)^{1/2}.$$

Proof. For each $0 < r < R$, we set $f(r) = \mathcal{H}^{n-1}(A \cap \rho^{-1}\{r\})$ and use Federer's coarea formula (§ 3.1.2(iv)) to estimate

$$\mathcal{L}^n(A) = \int_A J_1 \rho d\mathcal{L}^n = \int_0^R f(r) dr$$

and

$$\int_A \rho d\mathcal{L}^n = \int_0^R rf(r) dr.$$

The assertion of Step 2 now follows from Step 1.

Step 3. For $0 < R < \infty$, we note that

$$\mathcal{L}^n(A \cap \rho^{-1}(0, R)) \leq 2^{1/2} (\sup \{ \mathcal{H}^{n-1}(A \cap \rho^{-1}\{r\}) : 0 < r < R \})^{1/2} (\Delta t)^{1/2} E^{1/2}$$

as a consequence of Step 2, while clearly

$$R \mathcal{L}^n(A \cap \rho^{-1}[R, \infty)) \leq \int_{A \cap \rho^{-1}[R, \infty)} \rho d\mathcal{L}^n$$

so that $\mathcal{L}^n(A \cap \rho^{-1}[R, \infty)) \leq (\Delta t/R)E$. The proposition follows. \square

PROPOSITION 4.2 (a volume-distance inequality when the zero set has density ratios bounded from below). *Suppose that C is a compact subset of \mathbb{R}^n and $\rho(x) = \text{dist}(x, C)$ for each x . Suppose also that A is an \mathcal{L}^n measurable subset of \mathbb{R}^n and that $\delta, \theta, \Delta t$, and E are positive numbers such that*

$$\frac{1}{\Delta t} \int_A \rho \, d\mathcal{L}^n \leq E$$

and

$$\mathcal{H}^{n-1}(C \cap \mathbb{B}^n(p, r)) \geq \theta r^{n-1}.$$

whenever p lies in C and $0 < r \leq \delta$. Then, for each $\delta < R < \infty$,

$$\mathcal{L}^n(A \sim C) \leq \left[2\Gamma \left(\frac{R}{\delta} \right)^{n-1} \mathcal{H}^{n-1}(C) \right]^{1/2} (\Delta t)^{1/2} E^{1/2} + \frac{\Delta t}{R} E.$$

Here

$$\Gamma = 2^{2n+1} n \alpha(n) \beta(n) / \theta,$$

$\alpha(n)$ is the volume of the unit n ball in \mathbb{R}^n , and $\beta(n)$ is the constant of the Besicovitch-Federer covering theorem in \mathbb{R}^n .

Proof of Proposition 4.2. Given Proposition 4.1, we prove this proposition by showing that

$$\sup_{0 < r < R} \mathcal{H}^{n-1}(\rho^{-1}\{r\}) \leq \Gamma \left(\frac{R}{\delta} \right)^{n-1} \mathcal{H}^{n-1}(C).$$

We prove this in three steps, shown below. In them, we use Σ rather than C , because sometimes we are dealing with only a portion of C or a scaled and translated portion of it. Similarly, we use σ rather than ρ to denote distance to Σ . (The proposition will be applied with $C = \text{spt } \partial[K_j(k\Delta t_j)]$ in Theorem 4.5.)

Step 1 (a divergence theorem estimate between level sets of a distance function). Suppose that σ is the distance function to some compact subset Σ of \mathbb{R}^n . Then, for almost every $0 < R < S < \infty$,

$$\mathcal{H}^{n-1}(\sigma^{-1}\{S\}) - \mathcal{H}^{n-1}(\sigma^{-1}\{R\}) \leq \frac{n-1}{R} \mathcal{L}^n(\sigma^{-1}(R, S)).$$

Proof. We would like to prove Step 1 via the intermediate relations

$$\int_{\sigma^{-1}\{S\}} \nabla \sigma \cdot \mathbf{n} \, d\mathcal{H}^{n-1} - \int_{\sigma^{-1}\{R\}} \nabla \sigma \cdot \mathbf{n} \, d\mathcal{H}^{n-1} = \int_{\sigma^{-1}(R, S)} \Delta \sigma \, d\mathcal{L}^n \leq \frac{n-1}{R} \int_{\sigma^{-1}(R, S)} 1 \, d\mathcal{L}^n$$

with the equality following from the divergence theorem (with $\mathbf{n} = \mathbf{n}_{\sigma^{-1}(R, S)}$), and the inequality arising from the facts that the Laplacian of the distance function from a point is $(n-1)/R$ and that the infimum of two functions has its Laplacian less than the supremum of the two Laplacians (with the Laplacian along the crease where the two functions are equal being $-\infty$). However, σ need not be smooth enough to do this directly (unless C consists of just a couple of points). We must therefore work harder.

Suppose that Ω is an open subset of \mathbb{R}^n , that $f, g : \Omega \rightarrow \mathbb{R}$ are infinitely smooth, and that $\nabla(f-g)$ does not vanish on the smooth submanifold $\Gamma = \{x; f(x) = g(x)\}$; we set $A = \Omega \cap \{x; f(x) < g(x)\}$. We check that the distribution Laplacian of the function $h = \inf\{f, g\}$ is associated with the measure

$$\mathcal{L}^n L\Omega \wedge \Delta h + \mathcal{H}^{n-1} \llcorner \Gamma \wedge (\mathbf{n}_A \cdot \nabla g - \mathbf{n}_A \cdot \nabla f)$$

and that $\mathbf{n}_A \cdot \nabla g - \mathbf{n}_A \cdot \nabla f$ is negative; here Δh is set equal to zero on Γ . By saying this,

we mean, in particular, that

$$\int_{\Omega} \Delta \varphi h \, d\mathcal{L}^n = \int_{\Omega} \varphi \Delta h \, d\mathcal{L}^n + \int_{\Gamma} \varphi (\mathbf{n}_A \cdot \nabla g - \mathbf{n}_A \cdot \nabla f) \, d\mathcal{H}^{n-1}$$

for every test function φ with compact support in Ω . It follows that, if M is positive and $\Delta f \leq M$ and $\Delta g \leq M$, then

$$\int_{\Omega} \Delta \varphi h \, d\mathcal{L}^n \leq M \int_{\Omega} \varphi \, d\mathcal{L}^n$$

for every nonnegative test function φ with compact support in Ω . More generally, if f and g are summable functions with

$$\int_{\Omega} \Delta \varphi f \, d\mathcal{L}^n \leq M \int_{\Omega} \varphi \, d\mathcal{L}^n \quad \text{and} \quad \int_{\Omega} \Delta \varphi g \, d\mathcal{L}^n \leq M \int_{\Omega} \varphi \, d\mathcal{L}^n$$

for every nonnegative test function φ with compact support in Ω , then

$$\int_{\Omega} \Delta \varphi \inf \{f, g\} \, d\mathcal{L}^n \leq M \int_{\Omega} \varphi \, d\mathcal{L}^n$$

for such φ . Indeed, regularization and the Morse–Sard–Federer theorem reduce the estimate to the case in which f and g are smooth as above. The Laplacian of the radius function r is computed to equal $(n-1)/r$ for $r \neq 0$. Since σ is the infimum of a countable set of nonnegative radius functions centered at points dense in Σ , we use Lebesgue’s dominated convergence theorem to conclude that, whenever $0 < R < S < \infty$ and $\Omega = \sigma^{-1}(R, S)$, then

$$\int_{\Omega} \Delta \varphi \sigma \, d\mathcal{L}^n \leq \frac{n-1}{R} \int_{\Omega} \varphi \, d\mathcal{L}^n$$

for every nonnegative test function φ with compact support in Ω . Suppose that Ψ_{ε} is a spherically symmetric smoothing function with support contained in $\mathbb{B}^n(0, \varepsilon)$ and $\sigma_{\varepsilon} = \sigma * \Psi_{\varepsilon}$ is a smoothing of σ and $\nabla \sigma_{\varepsilon}$ is the smoothing of $\nabla \sigma$, equivalently, the gradient of σ_{ε} . Then, as is well known, $\sigma_{\varepsilon} \rightarrow \sigma$ uniformly as $\varepsilon \downarrow 0$ and $\nabla \sigma_{\varepsilon} \rightarrow \nabla \sigma$ almost everywhere as $\varepsilon \downarrow 0$. We use Federer’s coarea formula (§ 3.1.2(iv)) to confirm that, for almost every R and S , our Ω will be a set having finite perimeter $\partial\Omega$, which is \mathcal{H}^{n-1} almost equal to $\sigma^{-1}\{S\} \cup \sigma^{-1}\{R\}$, while, for \mathcal{H}^{n-1} almost every p in $\sigma^{-1}\{S\}$, $\mathbf{n}_{\Omega}(p) = \nabla \sigma(p)$, and for \mathcal{H}^{n-1} almost every p in $\sigma^{-1}\{R\}$, $\mathbf{n}_{\Omega}(p) = -\nabla \sigma(p)$. Similarly, for almost every R and S ,

$$\int_{\partial\Omega} |\nabla \sigma_{\varepsilon} - \nabla \sigma| \, d\mathcal{H}^{n-1} \rightarrow 0 \quad \text{as } \varepsilon \downarrow 0.$$

For almost every R and S , we can let χ_{Ω} denote the characteristic function of Ω and use the divergence theorem to infer that

$$\begin{aligned} \int_{\Omega} \Delta \sigma_{\varepsilon} \, d\mathcal{L}^n &= \int_{\partial\Omega} \nabla \sigma_{\varepsilon} \cdot \mathbf{n}_{\Omega} \, d\mathcal{H}^{n-1} \\ &= \int_{\sigma^{-1}\{S\}} \nabla \sigma \cdot \nabla \sigma \, d\mathcal{H}^{n-1} + \int_{\sigma^{-1}\{S\}} (\nabla \sigma_{\varepsilon} - \nabla \sigma) \cdot \mathbf{n}_{\Omega} \, d\mathcal{H}^{n-1} \\ &\quad - \int_{\sigma^{-1}\{R\}} \nabla \sigma \cdot \nabla \sigma \, d\mathcal{H}^{n-1} - \int_{\sigma^{-1}\{R\}} (\nabla \sigma_{\varepsilon} - \nabla \sigma) \cdot \mathbf{n}_{\Omega} \, d\mathcal{H}^{n-1} \\ &= \mathcal{H}^{n-1}(\sigma^{-1}\{S\}) - \mathcal{H}^{n-1}(\sigma^{-1}\{R\}) \\ &\quad + \int_{\sigma^{-1}\{S\}} (\nabla \sigma_{\varepsilon} - \nabla \sigma) \cdot \mathbf{n}_{\Omega} \, d\mathcal{H}^{n-1} - \int_{\sigma^{-1}\{R\}} (\nabla \sigma_{\varepsilon} - \nabla \sigma) \cdot \mathbf{n}_{\Omega} \, d\mathcal{H}^{n-1}, \end{aligned}$$

while

$$\begin{aligned}
 \int_{\Omega} \Delta \sigma_{\varepsilon} d\mathcal{L}^n &= \lim_{\delta \downarrow 0} \int_{\mathbb{R}^n} (\chi_{\Omega} * \Phi_{\delta}) \Delta \sigma_{\varepsilon} d\mathcal{L}^n \\
 &= \lim_{\delta \downarrow 0} \int_{\mathbb{R}^n} \Delta(\chi_{\Omega} * \Phi_{\delta}) \sigma_{\varepsilon} d\mathcal{L}^n \\
 &= \lim_{\delta \downarrow 0} \int_{\mathbb{R}^n} \Delta(\chi_{\Omega} * \Phi_{\delta} * \Phi_{\varepsilon}) \sigma d\mathcal{L}^n \\
 &\leq \lim_{\delta \downarrow 0} \frac{n-1}{R-\delta-\varepsilon} \int_{\mathbb{R}^n} (\chi_{\Omega} * \Phi_{\delta} * \Phi_{\varepsilon}) d\mathcal{L}^n \\
 &= \frac{n-1}{R-\varepsilon} \int_{\mathbb{R}^n} (\chi_{\Omega} * \Phi_{\varepsilon}) d\mathcal{L}^n.
 \end{aligned}$$

Our asserted inequality follows by combining these two inequalities and letting $\varepsilon \downarrow 0$.

Step 2 (an estimate on the size of level sets of a distance function inside balls). Suppose that $\sigma: \mathbb{R}^n \rightarrow \mathbb{R}^+$ gives the distance to some closed subset Σ of \mathbb{R}^n . Then

$$\mathcal{H}^{n-1}(\sigma^{-1}\{1\} \cap \mathbb{B}^n(0, 2)) \leq 2^{2n+1} n\alpha(n).$$

Proof. Replacing Σ by $\Sigma \cap \mathbb{B}^n(0, 3)$, if necessary we assume without loss of generality that $\Sigma \subset \mathbb{B}^n(0, 3)$. Federer's coarea formula (§ 3.1.2(iv)) implies the existence of $\frac{1}{2} < R < 1$ such that $\mathcal{H}^{n-1}(\sigma^{-1}\{R\}) \leq 2\mathcal{L}^n(\sigma^{-1}(\frac{1}{2}, 1))$. Combining this with Step 1 we obtain

$$\begin{aligned}
 \mathcal{H}^{n-1}(\sigma^{-1}\{1\}) &\leq \mathcal{H}^{n-1}(\sigma^{-1}\{R\}) + 2(n-1)\mathcal{L}^n(\sigma^{-1}(R, 1)) \\
 &\leq 2n\mathcal{L}^n(\sigma^{-1}(\tfrac{1}{2}, 1)) \\
 &\leq 2n\mathcal{L}^n(\mathbb{B}^n(0, 4)) \\
 &= 2^{2n+1} n\alpha(n).
 \end{aligned}$$

Step 3. For almost every $0 < r < \infty$,

$$\mathcal{H}^{n-1}(\rho^{-1}\{r\}) \leq \begin{cases} \Gamma \mathcal{H}^{n-1}(C) & \text{in case } 0 < r < \delta, \\ \Gamma(r/\delta)^{n-1} \mathcal{H}^{n-1}(C) & \text{in case } \delta \leq r < \infty, \end{cases}$$

where $\Gamma = 2^{2n+1} n\alpha(n)\beta(n)/\theta$ as above.

Proof. Fix $0 < r < \infty$ and consider balls $\mathbb{B}^n(p, r)$ corresponding to all points p in C . Then, according to our hypothesis about density ratio, for each p and r we have

$$r^{n-1} \leq \frac{1}{\theta} \mathcal{H}^{n-1}(C \cap \mathbb{B}^n(p, r)) \quad \text{in case } 0 < r < \delta$$

and

$$r^{n-1} = \left(\frac{r}{\delta}\right)^{n-1} (\delta)^{n-1} \leq \left(\frac{r}{\delta}\right)^{n-1} \frac{1}{\theta} \mathcal{H}^{n-1}(C \cap \mathbb{B}^n(p, \delta)) \quad \text{for } \delta \leq r < \infty.$$

The Besicovitch-Federer covering theorem (or a direct argument in this case) guarantees the existence of points p_1, p_2, p_3, \dots in C such that C is contained in the union of the $\mathbb{B}^n(p_i, r)$'s and no point in \mathbb{R}^n is contained in more than $\beta(n)$ of the

$\mathbb{B}^n(p_i, r)$'s. Since C is bounded, the number of these balls is finite. Clearly, $\rho^{-1}\{r\}$ is contained in the union of the $\mathbb{B}^n(p_i, 2r)$'s, since each point in $\rho^{-1}\{r\}$ is distance r from some point in C and that point is contained in one of the $\mathbb{B}^n(p_i, r)$'s. Hence, we can scale Step 2, above, by the homothetic factor r in an obvious way to estimate

$$\mathcal{H}^{n-1}(\rho^{-1}\{r\}) \leq \sum_i \mathcal{H}^{n-1}(\rho^{-1}\{r\} \cap \mathbb{B}^n(p_i, 2r)) \leq \sum_i 2^{2n+1} n\alpha(n) r^{n-1}.$$

We abbreviate $\tau = 2^{2n+1} n\alpha(n)/\theta$. In case $0 < r < \delta$, we further estimate

$$\mathcal{H}^{n-1}(\rho^{-1}\{r\}) \leq \tau \sum_i \mathcal{H}^{n-1}(C \cap \mathbb{B}^n(p_i, r)) \leq \tau \beta(n) \mathcal{H}^{n-1}(C).$$

In case $\delta \leq r < \infty$, we similarly estimate

$$\mathcal{H}^{n-1}(\rho^{-1}\{r\}) \leq \tau \left(\frac{r}{\delta}\right)^{n-1} \sum_i \mathcal{H}^{n-1}(C \cap \mathbb{B}^n(p_i, \delta)) \leq \tau \left(\frac{r}{\delta}\right)^{n-1} \beta(n) \mathcal{H}^{n-1}(C).$$

The assertion of Step 3 follows.

The assertion of the proposition follows from Step 3 and Proposition 4.1, as noted initially. \square

PROPOSITION 4.3 (an inequality for sums). *Suppose that $A_i, E_i, \Delta t_i$ are positive numbers for $i = 1, \dots, N$ and*

$$A = \sum_{i=1}^N A_i, \quad E = \sum_{i=1}^N E_i, \quad \Delta t = \sum_{i=1}^N \Delta t_i \leq 1.$$

Suppose also that, for each $i = 1, \dots, N$ and each $R > \Delta t_i$,

$$A_i \leq \Gamma \left(\frac{R}{\Delta t_i}\right)^{(n-1)/2} (\Delta t_i)^{1/2} E_i^{1/2} + \frac{\Delta t_i}{R} E_i$$

for some positive constant Γ . Then

$$A \leq (\Gamma E^{1/2} + E)(\Delta t)^{1/(n+1)}.$$

Proof. For each i , set $R_i = \Delta t_i / \Delta t^{1/(n+1)}$ and use Schwarz's inequality to estimate

$$\begin{aligned} A &\leq \sum_{i=1}^N \Gamma \left(\frac{R_i}{\Delta t_i}\right)^{(n-1)/2} (\Delta t_i)^{1/2} E_i^{1/2} + \frac{\Delta t_i}{R_i} E_i \\ &= \sum_{i=1}^N \Gamma (\Delta t)^{(1-n)/2(n+1)} (\Delta t_i)^{1/2} E_i^{1/2} + (\Delta t)^{1/(n+1)} E_i \\ &\leq \Gamma (\Delta t)^{(1-n)/2(n+1)} (\Delta t)^{1/2} E^{1/2} + (\Delta t)^{1/(n+1)} E \\ &= (\Gamma E^{1/2} + E)(\Delta t)^{1/(n+1)}. \end{aligned}$$

\square

THEOREM 4.4 (a Hölder estimate for the discrete approximations to flat Φ curvature flows). *Assume*

- (a) $[K(0)]$ is an initial solid in \mathcal{K} ;
- (b) $\mathcal{L}^n(\text{spt } \partial[K(0)]) = 0$;
- (c) j is a fixed positive integer and $\Delta t_j = 2^{-j}$;
- (d) $K_j(0) = K(0)$;
- (e) For each nonnegative integer k , we choose $[K_j((k+1)\Delta t_j)] \in \mathcal{K}$ so that

$$\mathcal{E}([K_j(k\Delta t_j)], [K_j((k+1)\Delta t_j)], \Delta t_j) = \inf \{ \mathcal{E}([K_j(k\Delta t_j)], [L], \Delta t_j) : [L] \in \mathcal{K} \};$$

- (f) $\Delta t = N\Delta t_j$, with N a positive integer such that $\Delta t \leq 1$.

Then there is a $\Gamma < \infty$ (given explicitly in the proof) such that, for any nonnegative integer k ,

$$\mathbb{G}(\partial[K_j(k\Delta t_j)] - \partial[K_j((k+N)\Delta t_j)]) \leq \Gamma \Delta t^{1/(n+1)}.$$

Proof of Theorem 4.4. Let $D = \text{diam}(K(0)) + 1$ and $E = \Phi(\partial[K(0)])$. For each nonnegative integer k , let $E_k = \Phi(\partial[K_j(k\Delta t_j)]) - \Phi(\partial[K_j((k+1)\Delta t_j)])$.

Step 1. For each k ,

$$\frac{1}{\Delta t_j} \int_{x \in K_j(k\Delta t_j) \Delta K_j((k+1)\Delta t_j)} \text{dist}((x, \partial K_j(k\Delta t_j))) d\mathcal{L}^n x \leq E_k.$$

Proof. This is a statement of the fact that $[K_j(l\Delta t_j)]$ is among the candidates for $[K_j((l+1)\Delta t_j)]$ in the minimization, and $\mathcal{E}([K], [K], c) = \Phi(\partial[K])$ for any solid K and any $c > 0$.

Step 2. For each k , $\mathcal{H}^{n-1}(\partial K_j(k\Delta t_j)) \leq E/\Phi_0$.

Proof. The volume integral part of Step 1 cannot be negative, so $E_l \geq 0$, for each $0 \leq l$. For each k , we add these inequalities for $0 \leq l \leq k-1$ to conclude that $\Phi(\partial[K_j(0)]) - \Phi(\partial[K_j(k\Delta t_j)]) > 0$. Step 2 follows.

Step 3. Let $\Gamma_0 = (2^{2n+1} n \alpha(n) \beta(n)) / \theta$ and $\Gamma_1 = (2\Gamma_0 E / \Phi_0)$. Here θ is the lower bound to the density ratios of § 3.5, and $\alpha(n)$ and $\beta(n)$ are as in § 2.1.1. Then

$$\begin{aligned} & \mathbb{G}(\partial[K_j(k\Delta t_j)] - \partial[K_j((k+1)\Delta t_j)]) \\ &= \mathcal{L}^n(K_j(k\Delta t_j) \triangle K_j((k+1)\Delta t_j)) \\ &\leq \left[2\Gamma_0 \left(\frac{R}{\Delta t_j} \right)^{n-1} \mathcal{H}^{n-1}(\partial K_j(k\Delta t_j)) \right]^{1/2} (\Delta t_j)^{1/2} E_k^{1/2} \\ &\leq \Gamma_1 \left(\frac{R}{\Delta t_j} \right)^{(n-1)/2} (\Delta t_j)^{1/2} + \frac{\Delta t_j}{R} E_k \end{aligned}$$

for each R with $\Delta t_j < R < \infty$.

Proof. The equality is a definition. The first inequality follows by applying Proposition 4.2 with C, δ, θ, E, A replaced there by $\text{spt } \partial[K_j(k\Delta t_j)], \Delta t_j, \theta, E_k, K_j(k\Delta t_j) \triangle K_j((k+1)\Delta t_j)$, respectively. (The hypotheses of Proposition 4.2 are satisfied, according to Step 1 and § 3.5.) The second inequality follows from the definition of Γ_1 and Step 2.

Step 4. Let $\Gamma = \Gamma_1 E^{1/2} + E$. Then, for any nonnegative integer k ,

$$\mathbb{G}(\partial[K_j(k\Delta t_j)] - \partial[K_j((k+N)\Delta t_j)]) \leq \Gamma \Delta t^{1/(n+1)}.$$

Proof. By the triangle inequality,

$$\mathbb{G}(\partial[K_j(k\Delta t_j)] - \partial[K_j((k+N)\Delta t_j)]) \leq \sum_{i=1}^N \mathbb{G}(\partial[K_j(k-1+i\Delta t_j)] - \partial[K_j((k+i)\Delta t_j)]).$$

We now apply Proposition 4.3, with $A, A_i, E, E_i, \Delta t, \Delta t_i, R, \Gamma$ replaced there by $\mathbb{G}(\partial[K_j(k\Delta t_j)] - \partial[K_j(k\Delta t_j + \Delta t)]), \mathbb{G}(\partial[K_j(k-1+i\Delta t_j)] - \partial[K_j((k+i)\Delta t_j)]), E', E_i, \Delta t, \Delta t_j, R, \Gamma_1$, respectively, where $E' = \sum_i E_{k-1+i} \leq E$. The hypotheses of Proposition 4.3 are satisfied according to Step 3. \square

THEOREM 4.5 (existence and Hölder continuity of flat Φ curvature flows). *Let $[K(0)]$ be an initial solid in \mathcal{K} for which $\mathcal{L}^n(\text{spt } \partial[K(0)]) = 0$ and suppose that approximate flat Φ curvature flows $[K_j(\cdot)]: \mathbb{R}^+ \rightarrow \mathcal{K}$ have been constructed as in § 2.6. Then there is a subsequence $j(1), j(2), j(3), \dots$ of $1, 2, 3, \dots$ and a flat Φ curvature flow $[K(\cdot)]: \mathbb{R}^+ \rightarrow \mathcal{K}$ such that*

$$\mathbb{G}(\partial[K_{j(i)}(\cdot)] - \partial[K(\cdot)]) \rightarrow 0 \quad \text{as } i \rightarrow \infty$$

locally uniformly. Furthermore, whenever $0 \leq s < t \leq s+1 < \infty$, then

$$\mathbb{G}[\partial[K(s)] - \partial[K(t)]] \leq \Gamma \Phi[\partial[K(0)]] |s - t|^{1/(n+1)}$$

for any such limit; here

$$\Gamma = 1 + \frac{2\Gamma_0}{\Phi_0},$$

where Γ_0 is the constant of Theorem 4.4, depending on the dimension n , the upper and lower bounds Φ^0 , Φ_0 of the surface energy density, and the diameter of $K(0)$.

Proof of Theorem 4.5. As stated in the Introduction, we infer from the existence theorem for \mathcal{E} minimizers (§ 3.2) that all the $K_j(t)$'s lie within the convex hull of the closure of $K(0)$ and have boundaries with masses dominated by E/Φ_0 ; hence the compactness theorem for integral currents (§ 3.1.5), together with Cantor's diagonal process, guarantees the existence of a subsequence $j(1), j(2), j(3), \dots$ of $1, 2, 3, \dots$ and a function $[K_0(\cdot)]$ such that $[K_{j(i)}(t)] \rightarrow [K(t)]$ as $i \rightarrow \infty$ for each fixed positive dyadic fractional number t (i.e., each $t = k2^{-m}$ for some integers k, m). The above assertions, for all positive s, t , now follow readily from the uniform Hölder continuity estimate of Theorem 4.4. The continuity at zero follows from the observation that, if $\lim_{\Delta t \downarrow 0} \mathcal{L}^n(K(0) \triangle L_{\Delta t})$ were not zero, then the \mathcal{E} energies of the minimizers would go to infinity, which they do not, since K_0 itself is a candidate in each case. \square

Remark. Stefan Luckhaus told us on July 23, 1992 that, in a forthcoming preprint [LS], he and Sturzenhecker will give a different argument for showing the Hölder continuity of flat curvature flows for the area integrand. Their estimate has exponent $\frac{1}{2}$, in contrast with our exponent $1/(n+1)$, above.

Note added in proof. A suggestion by Thomas Ilmanen in August 1992 has led us to an argument showing Hölder exponent $\frac{1}{2}$ for all Φ .

5. Wulff shapes and \mathcal{E} minimizers.

5.1. Wulff shapes. The *Wulff shape* for our surface energy function Φ is the current $[\mathcal{W}]$ associated with the convex body

$$\mathcal{W} =: \mathbb{R}^n \cap \{x: x \cdot u \leq \Phi(u) \text{ for each unit vector } u \in \mathbb{R}^n\}$$

by giving it positive orientation. Also, we denote by \mathcal{W}^* the central inversion of \mathcal{W} ; when it is used as a current (e.g., in evaluating surface energy), it is typically given negative orientation and thus written $-\mathcal{W}^*$.

By a *scaled Wulff shape* for Φ , we mean a solid $[W_R]$ (for some positive R) that is a homothetic image of $[\mathcal{W}]$, the scale chosen so that the result has volume R^n . A basic property [T1] of any scaled Wulff shape $[W_R]$ is that

$$\Phi(\partial[W_R]) \leq \Phi(\partial[L])$$

whenever $[L] \in \mathcal{H}$ with $\mathcal{L}^n(L) = \mathcal{L}^n(W_R) = R^n$. In our constructions, we will use scaled Wulff shapes $[W_R]$ in constructing comparison currents in \mathcal{E} minimizations and use the underlying sets W_R in constructing barriers by set addition. Similarly, we will use *scaled inverted Wulff shapes* $-\mathcal{W}_R^*$, which have the property that

$$\Phi(-\partial[W_R^*]) \leq \Phi(-\partial[L])$$

whenever $[L] \in \mathcal{H}$ with $\mathcal{L}^n(L) = \mathcal{L}^n(W_R) = R^n$. Scaled Wulff shapes and scaled inverted Wulff shapes for our surface energy Φ are analogous to soap bubbles for surface area in the sense that, for the volume enclosed (appropriately positively or negatively oriented), the surface energy attains its minimum value. (We think of the Wulff shape

as being the equilibrium shape of a crystal within another substance or vapor and the inverted Wulff shape as being the equilibrium shape of a hole in the crystal filled with the other substance.)

This extremal property, together with the ability to adjust volumes by homothetic expansions or contractions, enables us to obtain useful estimates, in terms of volume differences, on the surface energy of a Wulff shape with a small piece added or removed. Since our flat Φ curvature flows are limits of minimizers of \mathcal{E} , there is the potential for useful estimates if distance to the previous boundary, say ∂K_0 , can be controlled. Such distance control is obtained if we assume either that ∂K_0 lies inside a larger Wulff shape or outside a smaller one. When things are sorted out, we can conclude that flat Φ curvature flows that start outside an appropriate initial Wulff shape will remain outside a suitably shrinking Wulff shape, while those starting inside an appropriate Wulff shape remain inside Wulff shapes shrinking at a somewhat different rate. We also use scaled inverted Wulff shapes when outside the crystal solid. *Caution:* All the assertions in this chapter are based on Wulff shapes shrinking from initial time $t = 0$ and are not known to hold when starting at a later time $t_0 > 0$ unless all the $\partial K_j(t_0)$'s ($j = 1, 2, 3, \dots$) satisfy the starting assumptions (and not only $\partial L(t_0)$).

We set

$$w_* =: \sup \{r: \mathbb{B}^n(0, r) \subset W_1\} \leq w^* =: \inf \{r: \mathbb{B}^n(0, r) \supset W_1\}.$$

If $0 < R < 1$, then the following statements hold:

- (i) Each point in $\mathbb{R}^n \sim W_1$ is distance at least $(1 - R)w_*$ from W_R ;
- (ii) Each point in W_1 is distance no more than $(1 - R)w^*$ from W_R .

To see (i), we note that the infimum of the distances between points in the closure of ∂W_1 and in the closure of ∂W_R is attained and necessarily the support planes to W_1 and W_R are unique and parallel. The homothety carrying W_1 to W_R carries one of these planes into the other. The infimum of distances between such planes is clearly $(1 - R)w_*$. If $p \in W_1$, then $Rp \in W_R$ with $|p - Rp| = (1 - R)|p| \leq (1 - R)w^*$; (ii) follows.

PROPOSITION 5.2 (estimates on perturbations of a Wulff shape). *Let $[W] \in \mathcal{H}$ be a scaled Wulff shape for Φ and suppose that $[U], [V]$ are also solids of positive mass in \mathcal{H} . Then*

- (1) $(\mathcal{L}^n(W)/\mathcal{L}^n(U))^{(n-1)/n} \Phi(\partial[U]) \geq \Phi(\partial[W])$;
- (2) *If $V \subset W$ with $2\mathcal{L}^n(V) \leq \mathcal{L}^n(W)$, then*

$$\begin{aligned} \Phi(\partial[W]) - \Phi(\partial[W \sim V]) &\leq \left(1 - \left[1 - \frac{\mathcal{L}^n(V)}{\mathcal{L}^n(W)}\right]^{(n-1)/n}\right) \Phi(\partial[W]) \\ &\leq \Phi(\partial[W]) \left[\frac{n-1}{n} \left(\frac{\mathcal{L}^n(V)}{\mathcal{L}^n(W)}\right) + \frac{2^{1/n}(n-1)}{n^2} \left(\frac{\mathcal{L}^n(V)}{\mathcal{L}^n(W)}\right)^2 \right]; \end{aligned}$$

- (3) *If $V \subset \mathbb{R}^n \sim W$, then*

$$\begin{aligned} \Phi(\partial[W \cup V]) - \Phi(\partial[W]) &\geq \left(\left[1 + \frac{\mathcal{L}^n(V)}{\mathcal{L}^n(W)}\right]^{(n-1)/n} - 1\right) \Phi(\partial[W]) \\ &\geq \Phi(\partial[W]) \left[\frac{n-1}{n} \left(\frac{\mathcal{L}^n(V)}{\mathcal{L}^n(W)}\right) - \frac{n-1}{2n^2} \left(\frac{\mathcal{L}^n(V)}{\mathcal{L}^n(W)}\right)^2 \right]. \end{aligned}$$

Now, suppose additionally that $[K], [L]$ are solids of positive mass in \mathcal{H} ;

- (4) *If $L = K \cup W$ and $2\mathcal{L}^n(W \sim K) \leq \mathcal{L}^n(W)$, then*

$$\begin{aligned} \Phi(\partial[L]) - \Phi(\partial[K]) \\ \leq \Phi(\partial[W]) \left[\frac{n-1}{n} \left(\frac{\mathcal{L}^n(W \sim K)}{\mathcal{L}^n(W)}\right) + \frac{2^{1/n}(n-1)}{n^2} \left(\frac{\mathcal{L}^n(W \sim K)}{\mathcal{L}^n(W)}\right)^2 \right]; \end{aligned}$$

(5) If $L = K \cap W$, then

$$\begin{aligned} & \Phi(\partial[K]) - \Phi(\partial[L]) \\ & \geq \Phi(\partial[W]) \left[\frac{n-1}{n} \left(\frac{\mathcal{L}^n(K \sim W)}{\mathcal{L}^n(W)} \right) - \frac{n-1}{2n^2} \left(\frac{\mathcal{L}^n(K \sim W)}{\mathcal{L}^n(W)} \right)^2 \right]. \end{aligned}$$

Proof of Proposition 5.2. To establish the first assertion, we apply a homothetic transformation to U with scale factor $(\mathcal{L}^n(W)/\mathcal{L}^n(U))^{1/n}$ to adjust the volume of U to equal that of W . The Φ surface energy scales by factor $(\mathcal{L}^n(W)/\mathcal{L}^n(U))^{(n-1)/n}$. The left-hand inequalities of assertions (2) and (3) follow from (1). The right-hand inequalities of (2) and (3) respectively follow from the extended mean value theorem applied to the function $(1+x)^{(n-1)/n}$ in ranges $-\frac{1}{2} \leq x \leq 0$ and $0 \leq x < \infty$, respectively.

To establish (4), we set $V = W \sim K \subset W$ and use the last inequality in § 3.1.4 to check that

$$\begin{aligned} \Phi(\partial[L]) + \Phi(\partial[W \sim V]) &= \Phi(\partial[W \cup K]) + \Phi(\partial[W \cap K]) \\ &\leq \Phi(\partial[W]) + \Phi(\partial[K]); \end{aligned}$$

assertion (4) now follows from (2).

To establish (5), we set $V = K \sim W \subset \mathbb{R}^n \sim W$ and use the last inequality in § 3.1.4 to check that

$$\begin{aligned} \Phi(\partial[K]) + \Phi(\partial[W]) &\geq \Phi(\partial[K \cup W]) + \Phi(\partial[K \cap W]) \\ &= \Phi(\partial[W \cup V]) + \Phi(\partial[L]); \end{aligned}$$

assertion (5) now follows from (3). \square

PROPOSITION 5.3 (\mathcal{E} minimizers and Wulff shapes). *Suppose that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt . Suppose also that $0 < R < S < \infty$.*

(1) If $W_S \subset K_0$ and $0 < 2\mathcal{L}^n(W_R \sim K) \leq R^n$, then

$$\frac{S-R}{\Delta t} R \leq \frac{n-1}{n} \frac{\Phi(\partial[W_1])}{w_*} + \frac{2^{1/n}(n-1)}{n^2} \frac{\Phi(\partial[W_1])}{w_*} \frac{\mathcal{L}^n(W_R \sim K)}{R^n};$$

(2) If

$$R = \frac{S}{2} = \Delta t^{1/2} \left(\frac{2\Phi(\partial[W_1])}{w_*} \right)^{1/2},$$

then

$$\mathbb{R}^n \sim K \subset (\mathbb{R}^n \sim K_0) + W_S^* \quad (\text{set addition}).$$

(3) If $K_0 \subset W_S$ and $\mathcal{L}^n(K \sim W_R) > 0$, then

$$\frac{S-R}{\Delta t} R \geq \frac{n-1}{n} \frac{\Phi(\partial[W_1])}{w^*} - \frac{n-1}{2n^2} \frac{\Phi(\partial[W_1])}{w^*} \frac{\mathcal{L}^n(K \sim W_R)}{R^n}.$$

Proof. Suppose that $W_S \subset K_0$ and $2\mathcal{L}^n(W_R \sim K) \leq R^n = \mathcal{L}^n(W_R)$. As a comparison surface to our minimizer K , we set $L = K \cup W_R = K \cup (W_R \sim K)$ and check that $K_0 \triangle L \subset K_0 \triangle K$ with $K_0 \triangle K \sim K_0 \triangle L = W_R \sim K$. Since K is an \mathcal{E} minimizer, we conclude using statement 5.1(i) that

$$\begin{aligned} \Phi(\partial[L]) - \Phi(\partial[K]) &\geq \frac{1}{\Delta t} \int_{x \in (K_0 \triangle K) \sim (K_0 \triangle L)} \text{dist}(x, \partial K_0) d\mathcal{L}^n x \\ &\geq \frac{1}{\Delta t} \mathcal{L}^n(W_R \sim K) S \left(1 - \frac{R}{S} \right) w_*. \end{aligned}$$

Assertion (1) follows from conclusion (4) of Proposition 5.2.

To verify assertion (2), we check that our assumption implies that

$$\frac{R^2}{\Delta t} > \frac{n-1}{n} \frac{\Phi(\partial[W_1])}{w_*} + \frac{2^{1/n}(n-1)}{n^2} \frac{\Phi(\partial[W_1])}{w_*} \frac{1}{2}.$$

If the asserted inclusion did not hold, then the set $(\mathbb{R}^n \sim K) \sim [(\mathbb{R}^n \sim K_0) + W_S^*]$ would contain some point p . Applying a translation, if necessary, we assume without loss of generality that p is the origin in \mathbb{R}^n and hence infer that $W_S \subset K_0$ and $\mathcal{L}^n(W_R \sim K) > 0$. In this case, we additionally assert that $\mathcal{L}^n(W_R \sim K) \leq R^n/2$. If this inequality did not hold, i.e., if $\mathcal{L}^n(W_R \sim K) > R^n/2$, we could construct a comparison set $L = K \cup W_R$ as above and, as in the above proof of (1), confirm that

$$\Phi(\partial[L]) - \Phi(\partial[K]) \geq \frac{1}{\Delta t} \mathcal{L}^n(W_R \sim K) S \left(1 - \frac{R}{S}\right) w_* > \frac{w_* R^{n+1}}{2\Delta t}.$$

We use the last inequality in § 3.1.4 to further estimate that

$$\Phi(\partial[L]) \leq \Phi(\partial[K \cup W_R]) + \Phi(\partial[K \cap W_R]) \leq \Phi(\partial[K]) + \Phi(\partial[W_R]),$$

so that

$$\Phi(\partial[L]) - \Phi(\partial[K]) \leq R^{n-1} \Phi(\partial[W_1]),$$

and hence

$$\frac{w_* R^{n+1}}{2\Delta t} < R^{n-1} \Phi(\partial[W_1]),$$

which is false by our definition of R . Under our original assumption that (2) fails, we conclude that the hypotheses of (1) are satisfied, from which we infer that

$$\frac{S-R}{\Delta t} R \leq \frac{n-1}{n} \frac{\Phi(\partial[W_1])}{w_*} + \frac{2^{1/n}(n-1)}{n^2} \frac{\Phi(\partial[W_1])}{w_*} \frac{1}{2},$$

which is also false by our assumptions on R . This completes the proof of (2).

To verify assertion (3), we assume that $K_0 \subset W_S$. As a comparison surface to our minimizer K , we set $L = K \cap W_R$ so that $K_0 \triangle L \sim K_0 \triangle K = K \sim W_R$. Since K is an \mathcal{E} minimizer, we conclude using statement 5.1(ii) that

$$\begin{aligned} \Phi(\partial[K]) - \Phi(\partial[L]) &\leq \frac{1}{\Delta t} \int_{x \in (K_0 \triangle L) \sim (K_0 \triangle K)} \text{dist}(x, \partial K_0) d\mathcal{L}^n x \\ &\leq \frac{1}{\Delta t} \mathcal{L}^n(K \sim W_R) S \left(1 - \frac{R}{S}\right) w^*. \end{aligned}$$

Assertion (3) follows from part (5) of Proposition 5.2. \square

THEOREM 5.4 (flat Φ curvature flows and shrinking Wulff sets). *Assuming that $0 < S < \infty$ is given, we set*

$$\begin{aligned} S_0(t) &= (S^2 - c_0 t)^{1/2} \quad \text{for } 0 \leq t \leq \frac{S^2}{c_0}, \\ S^0(t) &= (S^2 - c^0 t)^{1/2} \quad \text{for } 0 \leq t \leq \frac{S^2}{c^0}; \end{aligned}$$

here

$$c_0 = \frac{2(n-1)}{n} \frac{\Phi(\partial[W_1])}{w_*}, \quad c^0 = c_0 \frac{w_*}{w^*}.$$

For later use, we set

$$c_1 = \frac{w_*}{16\Phi(\partial[W_1])}.$$

Assume that $\partial[K_j(\cdot)]:\mathbb{R}^+ \rightarrow \partial\mathcal{H}$ are approximate flat Φ curvature flows for $j=1, 2, 3, \dots$ associated with timesteps $\Delta t_j = 2^{-j}$. Assume also that $\partial[K(\cdot)]:\mathbb{R}^+ \rightarrow \partial\mathcal{H}$ is a flat Φ curvature flow that is the limit of a subsequence of the $\partial[K_j(\cdot)]$'s and that p is a point in \mathbb{R}^n .

Then the following are true:

- (1) Suppose that $\{p\} + W_S \subset K_j(k_0\Delta t_j)$ for some j and k_0 . If j is large enough so that $\Delta t_j < c_1^2 S^2$ and k is small enough so that $(k-1)\Delta t_j \leq S/2$, then $\{p\} + W_{S-(2c_0/S)k\Delta t_j} \subset K_j((k_0+k)\Delta t_j)$;
- (2) If $\{p\} + W_S \subset K(0)$, then $\{p\} + W_{S_0(t)} \subset K(t)$ for each $0 \leq t \leq S^2/c_0$;
- (3) Suppose that $\{p\} + W_S^* \subset \mathbb{R}^n \sim K(k_0\Delta t_j)$ for some j and k_0 . If j is large enough so that $\Delta t_j < c_1^2 S^2$ and k is small enough so that $(k-1)\Delta t_j \leq S/2$, then $\{p\} + W_{S-(2c_0/S)k\Delta t_j}^* \subset \mathbb{R}^n \sim K_j((k_0+k)\Delta t_j)$;
- (4) If $\{p\} + W_S^* \subset \mathbb{R}^n \sim K(0)$, then $\{p\} + W_{S_0(t)}^* \subset \mathbb{R}^n \sim K(t)$ for each $0 \leq t \leq S^2/c_0$;
- (5) If $K_j(k_0\Delta t_j) \subset \{p\} + W_S$ for some j and k_0 , then, for each $k=1, 2, 3, \dots$,

$$K_j((k_0+k)\Delta t_j) \subset \{p\} + W_{\sup\{S/2, S-(2c^0/S)k\Delta t_j\}};$$

- (6) If $K(0) \subset \{p\} + W_S$, then $K(t) \subset \{p\} + W_{S^0(t)}$ for each $0 \leq t \leq S^2/c^0$.

Proof.

Step 1. Whenever $0 < R < S < \infty$, we check that

$$W_S = W_R + W_{S-R} \quad \text{and} \quad \mathbb{R}^n \sim W_R = (\mathbb{R}^n \sim W_S) + W_{S-R}^*.$$

Step 2. Suppose that $0 < S < \infty$ and $[K_0] \in \mathcal{H}$ with $W_S \subset K_0$. For each

$$0 < \Delta t < \frac{S^2}{16} \left(\frac{w_*}{2\Phi(\partial[W_1])} \right)^2,$$

we choose $[K_{\Delta t}] \in \mathcal{H}$ such that

$$\mathcal{E}([K_0], [K_{\Delta t}], \Delta t) = \inf \{ \mathcal{E}([K_0], [L], \Delta t) : [L] \in \mathcal{H} \}.$$

We infer from assertion (2) of Proposition 5.3 and Step 1 that $W_{S/2} \sim K_{\Delta t} = \emptyset$ and set

$$R(\Delta t) = \sup \{ r : W_r \sim W_{\Delta t} = \emptyset \} \geq \frac{S}{2}.$$

We then infer from assertion (1) of Proposition 5.3 that

$$\frac{S-R(\Delta t)}{\Delta t} R(\Delta t) \leq \frac{n-1}{n} \frac{\Phi(\partial[W_1])}{w_*} = \frac{c_0}{2},$$

which implies, in addition, that

$$S-R(\Delta t) \leq \Delta t \frac{2(n-1)}{nS} \frac{\Phi(\partial[W_1])}{w_*}.$$

We infer conclusion (1) of Theorem 5.4 from this last formula. Suppose then that $f:[0, \Delta t] \rightarrow \mathbb{R}^+$ is a smooth decreasing function with $f(0) = S$ and, for each $0 < c < \Delta t$,

$$-(f^2)'(c) \geq c_0 + \frac{\Delta t}{S^2} \left(\frac{2(n-1)}{n} \frac{\Phi(\partial[W_1])}{w_*} \right)^2.$$

Then, according to the mean value theorem, there is $0 < c_* < \Delta t$ such that

$$\begin{aligned} -(f^2)'(c_*) &= \frac{f^2(0) - f^2(\Delta t)}{\Delta t} \\ &\geq c_0 + \frac{\Delta t}{S^2} \left(\frac{2(n-1)}{n} \frac{\Phi(\partial[W_1])}{w_*} \right)^2 \\ &\geq 2 \frac{S - R(\Delta t)}{\Delta t} R(\Delta t) + \frac{\Delta t}{S^2} \left(\frac{2(n-1)}{n} \frac{\Phi(\partial[W_1])}{w_*} \right)^2, \end{aligned}$$

which implies that

$$\begin{aligned} S^2 - f^2(\Delta t) &\geq (S - R(\Delta t))((S + R(\Delta t)) - (S - R(\Delta t))) \\ &\quad + \left(\frac{\Delta t}{S} \right)^2 \left(\frac{2(n-1)}{n} \frac{\Phi(\partial[W_1])}{w_*} \right)^2 \end{aligned}$$

so that

$$R(\Delta t)^2 \geq f^2(\Delta t) - (S - R(\Delta t))^2 + \left(\frac{\Delta t}{S} \right)^2 \left(\frac{2(n-1)}{n} \frac{\Phi(\partial[W_1])}{w_*} \right)^2 \geq f^2(\Delta t).$$

Since $S_0(\cdot)$ satisfies the equation $-(S_0^2(\cdot))' = c_0$, we infer conclusion (2) of the theorem.

Step 3. In an entirely straightforward way, we can reformulate the study of flat Φ curvature flows $[K(\cdot)]$ in the language of the complementary currents

$$[\mathbb{R}^n \sim K(\cdot)] = \mathbb{E}^n - [K(\cdot)] = \mathbf{t}(\mathbb{R}^n \sim K(\cdot), 1, \mathbf{e}_1 \wedge \dots \wedge \mathbf{e}_n).$$

The main observation is that, whenever $[K], [L] \in \mathcal{H}$, then

$$\Phi(\partial[L]) + \frac{1}{\Delta t} \int_{x \in K \triangle L} \text{dist}(x, \partial K) d\mathcal{L}^n x$$

equals

$$\Phi(-\partial[\mathbb{R}^n \sim L]) + \frac{1}{\Delta t} \int_{x \in (\mathbb{R}^n \sim K) \triangle (\mathbb{R}^n \sim L)} \text{dist}(x, \partial(\mathbb{R}^n \sim K)) d\mathcal{L}^n x$$

so that the \mathcal{E} minimizations in constructing approximations would be formally the same (since $\mathbf{n}_{\mathbb{R}^n \sim L} = -\mathbf{n}_L$); we must just use negative orientations for the evaluation of surface energies. Since the constructions used in proving conclusions (1) and (2) of Proposition 5.3 are entirely local, they admit a straightforward reformation when the assumption $W_S \subset K_0$ is replaced by the assumption $W_S^* \subset \mathbb{R}^n \sim K_0$. The arguments above used to prove conclusions (1) and (2) of Theorem 5.4 translate into proofs of conclusions (3) and (4) of the theorem, respectively.

Step 4. The proof of conclusions (5) and (6), above, are based on estimate (3) of Proposition 5.3 and are similar (but easier) than the proofs of conclusions (1) and (2), above. We leave them to the reader. \square

6. Hausdorff distance convergence and viscosity solutions. One of the key components in our study of flat Φ curvature flows is establishing their relationship to smooth Φ curvature flows. This is a substantial task because the discrete approximating flows are possibly of varying topological types and, in higher dimensions, can be partially singular. Examples show that, in mean curvature flow of some surfaces in space, singularities must develop at times before the final disappearance. The main intermediate steps between flat flows and smooth flows are viscosity Φ curvature flows.

In this chapter, we show in Theorem 6.2 that, if the supports of the currents in a flat Φ curvature flow vary continuously in the Hausdorff distance topology and if they are the uniform limits of the supports of the approximating flows in the same topology, then the supports of the flat Φ curvature flow constitute a viscosity Φ curvature flow. The smoothness of such a viscosity flow is discussed in § 2.11. It should be noted, however, that although the existence of a limit flat Φ curvature flow is proved in Theorem 4.5 (regardless of whether Φ is smooth and elliptic) the $\mathbb{H}\mathbb{D}$ continuity is *not* known to hold in general.

LEMMA 6.1 (comparison of weighted mean curvatures). *Suppose that Φ is smooth and elliptic and that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt , so that $B = \text{spt } \partial[K]$ is smooth almost everywhere (by conclusion (4) of Theorem 3.8). Suppose further that $h: \mathbb{R}^n \rightarrow \mathbb{R}$ is a twice continuously differentiable test function so that $\Sigma = \mathbb{R}^n \cap \{x: h(x) = 0\}$ is closed and $\Sigma^+ = \mathbb{R}^n \cap \{x: h(x) > 0\}$ is open. Suppose finally that $B \cap \Sigma^+$ is empty and $B \cap \Sigma$ consists of a single point p at which $\nabla h(p) \neq 0$ so that Σ is a smooth hypersurface near p . Then p is a regular point of B with a well-defined unit exterior normal vector $\mathbf{n}_K(p)$. We orient Σ near p to have orienting normal vector $\mathbf{n}_K(p)$. Then*

$$H_{\Phi, \partial K}(p) \cdot \nabla h(p) \leq H_{\Phi, \Sigma}(p) \cdot \nabla h(p) = \mathbf{n}(p) \cdot \nabla h(p) \text{ trace } (D^2\Phi(\mathbf{n}(p)) \circ D^2h(p)).$$

Proof. The one-sided barrier offered by Σ_+ near p is enough to guarantee that p is a regular point in accordance with part (1) of Theorem 3.10. The final assertion now follows from the definitions, since the situation is entirely classical. \square

THEOREM 6.2 (uniform $\mathbb{H}\mathbb{D}$ limits of flat flows are viscosity flows). *Suppose that Φ is elliptic and smooth and that $\partial[K_j(t)]$ are approximate curvature flows for $j = 1, 2, 3, \dots$ and $0 < t < \infty$ constructed as in § 2.6. Suppose also that $j(1), j(2), j(3), \dots$ is a subsequence of $1, 2, 3, \dots$ and $\partial[K(t)]$ is a limit flat Φ curvature flow for which*

$$\lim_{i \rightarrow \infty} \mathbb{G}(\partial[K_{j(i)}(t)] - \partial[K(t)]) = 0.$$

We set

$$A_j, A: \mathbb{R}^+ \rightarrow \mathcal{C}, \quad A_j(t) = \text{spt } \partial[K_j(t)], \quad A(t) = \text{spt } \partial[K(t)].$$

If A is $\mathbb{H}\mathbb{D}$ continuous for $a < t < b$ (some a and b) and

$$\lim_{i \rightarrow \infty} \mathbb{H}\mathbb{D}(A_{j(i)}(t), A(t)) = 0$$

uniformly for $a < t < b$, then $A(t)$ is a viscosity Φ curvature flow in (a, b) with respect to the orientation function $\sigma(x, t) = -1$ if $x \in \text{spt } [K(t)] \sim A(t)$ and $\sigma(x, t) = +1$ if $x \notin \text{spt } [K(t)]$.

Proof. Since A is $\mathbb{H}\mathbb{D}$ continuous in (a, b) , the set $C = \mathbb{R}^n \times (a, b) \cap \{(x, t): x \in A(t)\}$ is closed in $\mathbb{R}^n \times (a, b)$. To show that A is a viscosity Φ curvature flow (with respect to σ , above), we suppose that $g: \mathbb{R}^n \times (a, b) \rightarrow \mathbb{R}$ is some twice continuously differentiable test function so that $\Gamma =: \mathbb{R}^n \times (a, b) \cap \{(x, t): g(x, t) = 0\}$ is closed and $\Gamma^+ =: \mathbb{R}^n \times (a, b) \cap \{(x, t): g(x, t) > 0\}$ is open. We also suppose that $C \cap \Gamma^+$ is empty and that $C \cap \Gamma$ consists of a single point (p, s) at which $\nabla_{x,g}(p, s) \neq 0$. We must show that

$$\frac{\partial g}{\partial t}(p, s) \geq \text{trace} \left(D^2\Phi \left(\pm \frac{\nabla_{x,g}}{|\nabla_{x,g}|}(p, s) \right) \circ D^2_x g(p, s) \right);$$

here we take the $+$ sign if σ is positive in Γ^+ near (p, s) , and the $-$ sign otherwise. We now define $A_j^*(t) = \lim_{r \downarrow t} A_j(t) \cup \lim_{r \uparrow t} A_j(t)$ (taking the limit in the $\mathbb{H}\mathbb{D}$ topology) and set $C_j = \mathbb{R}^n \times (a, b) \cap \{(x, t): x \in A_j^*(t)\}$ so that each C_j is closed in $\mathbb{R}^n \times (a, b)$. By our uniqueness assumption on (p, s) , there will exist points $(p_i, s_i) \in C_{j(i)}$ converging

to (p, s) such that $\Gamma_i =: \mathbb{R}^n \times (a, b) \cap \{(x, t): g(x, t) = g(p_i, s_i)\}$ is closed, $\Gamma_i^+ =: \mathbb{R}^n \times (a, b) \cap \{(x, t): g(x, t) > g(p_i, s_i)\}$ is open, $C_i \cap \Gamma_i^+$ is empty, $(p_i, s_i) \in C_{j(i)} \cap \Gamma_i$, and $\nabla_x g(p_i, s_i) \neq 0$. For each i , we set $\Delta t_i = 2^{-j(i)}$ and estimate

$$\begin{aligned} \mathbf{n}_{K_i}(p) \cdot H_{\Phi, \partial K_{j(i)}(s_i)}(p_i) &= \pm \frac{\text{dist}(p_i, \partial K_{j(i)}(s_i))}{\Delta t_i} \\ &\approx \frac{g(p_i, s_i) - g(p_i, s_i - \Delta t_i)}{|\nabla_x g(p_i, s_i)| \Delta t_i} \\ &\approx \frac{\partial g}{\partial t}(p_i, s_i) \frac{1}{|\nabla_x g(p_i, s_i)|}. \end{aligned}$$

The theorem now follows from Lemma 6.1. \square

PROPOSITION 6.3. *Assume that our surface energy function Φ is smooth and elliptic. Suppose that $[K(0)] \in \mathcal{K}$ is an initial solid such that $\partial K(0)$ is a two times continuously differentiable $(n-1)$ -dimensional submanifold of \mathbb{R}^n without boundary. Then, for each $M < \infty$, there exists $\delta_0 = \delta_0(M) > 0$ with the following property.*

For $0 \leq t < \infty$, let $\partial[K_1(t)], \partial[K_2(t)], \partial[K_3(t)], \dots \in \mathcal{K}$ constitute approximate flows as in § 2.6, all starting with $\partial[K(0)]$. Suppose also that $j(1), j(2), j(3), \dots$ is a subsequence of $1, 2, 3, \dots$ and that $\partial[K(t)] \in \mathcal{K}$ gives a flat Φ curvature flow for which $\partial[K(t)] = \lim_{i \rightarrow \infty} \partial[K_{j(i)}(t)]$.

Suppose that $0 < \delta \leq \delta_0$ and, for each $0 < t < \delta$ and all sufficiently large j 's, that the set $\partial K_j(t)$ is a twice differentiable $(n-1)$ -dimensional submanifold of \mathbb{R}^n without boundary and that no principal curvature of any $\partial K_j(t)$ exceeds M at any point. Then

(1) *For each $j = 1, 2, 3, \dots$ and each positive integer k such that $k2^{-j} < \delta$ the set $\partial K_j(t)$ is a $2(k-1)$ times continuously differentiable $(n-1)$ -dimensional submanifold of \mathbb{R}^n ;*

(2) *We have convergence*

$$\mathbb{H}\mathbb{D}(\partial K_{j(i)}(t), \partial K(t)) \rightarrow 0$$

uniformly for $0 \leq t < \delta$ as $i \rightarrow \infty$;

(3) *$\partial K(t)$ is continuous in the $\mathbb{H}\mathbb{D}$ topology for $0 \leq t < \delta$;*

(4) *For each $0 < t_0 < \delta$ and each $p_0 \in \partial K(t)$, there exists a rotation $\theta: \mathbb{R}^n \rightarrow \mathbb{R}^n$ and a Lipschitz function $f: \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ and $r > 0$ such that the set*

$$\mathbb{B}^{n+1}((p_0, t_0), r) \cap [\{(x, f(x, t)), t: x \in \mathbb{R}^{n-1}, t \in \mathbb{R}\} \triangle \{(p, t): 0 < t < \delta, \theta(p) \in \partial K(t)\}]$$

is empty;

(5) *For $0 \leq t < \delta$, $\text{spt } \partial[K(t)] = \partial K(t)$ is a viscosity Φ curvature flow with respect to the natural orientation;*

(6) *For $0 \leq t < \delta$, $\partial[K(t)]$ is a smooth Φ curvature flow.*

Proof. According to [M2, Prop. 3.3], \mathcal{W} and \mathcal{W}^* (and hence W_R and W_R^* for any $R > 0$) are uniformly convex. Our assumption on the principal curvatures guarantees the existence of a single $S > 0$ such that, for all sufficiently large j 's, for each $0 \leq t < \delta$ and each $x \in \partial K_j(t)$, there are unique points $p \in K_j(t)$ and $q \in \mathbb{R}^n \sim K_j(t)$ such that

$$(\{p\} + W_S) \cap \partial K_j(t) = \{x\}, \quad (\{q\} + W_S^*) \cap \partial K_j(t) = \{x\}.$$

These estimates, together with assumptions (1) and (3) of Theorem 5.4, guarantee the existence of a constant $M_1 < \infty$ such that, for all sufficiently large j ,

$$\mathbb{H}\mathbb{D}(\partial K_j(t), \partial K_j(t + 2^{-j})) \leq M_1 \Delta t$$

whenever $0 \leq t < t + 2^{-j} < \delta$. Our assumption on principal curvatures implies that each of the $\partial K_j(t)$'s can be written locally as graph of a smooth function with very small first derivatives over a disk of fixed size. Our Hausdorff distance estimate implies that the coordinate systems used in the representation of $\partial K(0)$ as graphs can also be used for representing the $\partial K_j(t)$'s, provided that δ_0 is sufficiently small. Conclusion (1) of the proposition now follows from the higher regularity asserted in Theorem 3.8, since the signed distance to a C^k manifold is itself C^k near the manifold in accordance with [KP]. Our above remarks about coordinate system representations with small first derivatives together with our $\mathbb{H}\mathbb{D}$ estimate between hypersurfaces at different times, give the common Lipschitz bounds in space-time sufficient to imply conclusions (2)–(4). Conclusion (5) follows from conclusions (2) and (3), together with Theorem 6.2. Conclusion (6) follows from conclusion (4), together with [CW] and [W], as noted in § 2.11. \square

7. Existence and smoothness of flat Φ curvature flows. In this section, we establish the short time existence of classical Φ curvature flows, provided that our initial hypersurface is $C^{3+\alpha}$ and that Φ is smooth and elliptic. We also show that, for a normal velocity that is slightly faster or slightly slower than Φ curvature, we can similarly obtain flows for short times. The uniformity of these estimates implies that, if we assume a Φ curvature flow exists for a longer time, then there will also exist slightly faster or slower flows with nearby starting conditions for essentially the same time. These slightly perturbed flows then become barriers to flat Φ curvature flow and lead to a proof that flat Φ curvature flow coincides with smooth Φ curvature flow as long as the latter exists. If such a flow is written as a graph $x_n = f(x_1, x_2, \dots, x_{n-1}, t)$ then the various partial derivatives $\partial^2 f / \partial x_i \partial x_j$ and $\partial f / \partial t$ will be Hölder continuous and will satisfy our evolution equation in the ordinary way. We learned recently that Giga and Goto [GG] proved similar results for general curvature flows by a method similar to that used by Evans and Spruck [ES2, Chap 2]. Our proof is different from theirs.

THEOREM 7.1 (short time existence of smooth Φ curvature flows and flows with slightly perturbed velocities). *Suppose that our surface energy Φ is smooth and elliptic and that $[K(0)]$ is a solid in \mathcal{K} for which $\partial K(0)$ is a $C^{3+\alpha}$ hypersurface. Then there is a neighborhood U of $\partial K(0)$ in the $C^{2+\alpha}$ topology and a neighborhood V of zero in \mathbb{R} and a positive time T_0 such that, for each $\partial L(0)$ in U and each s in V , there is a smooth flow $\partial L(t)$ for $0 \leq t \leq T_0$ starting with $\partial L(0)$ such that the normal velocity of $\partial L(t)$ at point p in $\partial L(t)$ equals*

$$H_{\Phi, \partial L(t)}(p) + s \mathbf{n}_{L(t)}(p).$$

A more technical description of the statement of this result is developed during the course of the proof.

Proof. There are several parts of our proof.

(1) We introduce function space terminology with which to write evolving surfaces $\partial L(t)$ as graphs of real-valued functions $g(x, t)$ over $\mathcal{M} \times [0, T]$; here $\mathcal{M} = \partial K(0)$ is a fixed smooth domain manifold that is the boundary of a region $A = K(0)$ having unit exterior normal vector field $\mathbf{n} = \mathbf{n}_A : \mathcal{M} \rightarrow \mathbb{R}^n$ and $0 < T \leq 1$. At time t , we associate with the function $g(x, t)$ the surface

$$(7.1) \quad \{x + g(x, t)\mathbf{n}(x) : x \in \mathcal{M}\}.$$

We consider simultaneously Banach spaces of functions associated with all time intervals $[0, T]$, $0 < T \leq 1$ because we do not know in advance the interval of time for which we will be able to obtain solutions; in particular, the evolving surfaces might

quickly move out of the region in \mathbb{R}^n in which they can be represented smoothly in the form (7.1).

As a signed distance function defined for x in \mathbb{R}^n , we set $d(x) = \pm \text{dist}(x, \mathcal{M})$; here we take the negative sign if and only if x belongs to A .

(2) We will use the formula given in § 2.2 to derive the nonlinear parabolic PDE that a function $g(x, t)$ must satisfy for the surfaces given in (7.1) to constitute a smooth Φ curvature flow. This equation will be of the form

$$(7.2) \quad \frac{g_t}{|\mathbf{n} - (I - gD^2d)^{-1}\nabla g|} = G(x, g, Dg, D^2g).$$

Associated with this equation is the nonlinear operator F between Banach spaces of functions given by setting

$$(7.3) \quad F[g] = \left(\frac{g_t}{|\mathbf{n} - (I - gD^2d)^{-1}\nabla g|} - G(x, g, Dg, D^2g), g(\cdot, 0) \right).$$

Given f in $C^{2+\alpha}(M)$, finding a function $g(x, t)$ such that $F[g(x, t)] = (0, f(x))$ is equivalent to finding a smooth Φ curvature flow beginning with the surface

$$\{x + f(x)\mathbf{n}(x) : x \in \mathcal{M}\}.$$

Similarly, finding a function $g(x, t)$ such that $F[g(x, t)] = (s, f(x))$ is equivalent to finding a smooth flow beginning with the same surface but whose normal velocity equals the Φ curvature plus s times the unit normal vector to the surface at that time.

(3) We would like to apply the inverse function theorem [B, Thm. 3.15] to the function F to guarantee the existence of function $g(x, t)$ such that $F[g(x, t)] = (\varepsilon, f(x))$ for small ε and small initial function f . The hypotheses of this theorem involve the Fréchet derivatives

$$DF[g](h) = \lim_{s \rightarrow 0} \frac{F[g + sh] - F[g]}{s},$$

which, for each fixed g , gives a linear partial differential operator on the tangent (same) Banach space of functions h . We compute and exhibit these derivative $DF[g]$; by inspection, they depend continuously on g .

We must then show that each $DF[g]$ is an isomorphism. This is accomplished first by using the a priori estimates of [LSU, Thm. 10.1, Chap. 4] to obtain a mapping norm bound (using a stronger norm) on $DF[g]^{-1}$ in coordinate patches for those functions that are in the image. We use a maximum principle argument to infer that we can replace the stronger norm with the norm relevant for our problem.

We then show why the estimates on $DF[g]^{-1}$ are, in fact, independent of the time interval $[0, T]$ in our original problem, provided that $T \leq 1$.

Finally, we use the surjectivity criterion [GT, Thm. 5.2] that, in homotopies of bounded linear mappings between Banach spaces, either every map is surjective or none are. This reduces the surjectivity problem to finding a convenient linear parabolic operator, which we know to be surjective. We effectively choose a smooth metric on \mathcal{M} and use [LM, Thm. 5.2] to conclude that the heat operator in this new metric maps smooth functions surjectively onto smooth functions, and thus we can use the heat operator as that convenient linear parabolic operator.

(4) Knowing that the inverse mappings $DF[g]^{-1}$ are varying continuously and are uniformly bounded (for bounded g 's) regardless of the time interval in our domain, we must check that there is some $g_\delta(x, t)$ such that $F[g_\delta]$ is small. That being the case, the inverse function theorem implies that for all small (s, f) there will exist a unique

g such that $F[g] = (s, f)$ and that g depends continuously on s and f . The function $g_\delta(x, t)$ we take is constructed as follows. Let $G_\delta(x)$ be a slight smoothing of the function $G(x, 0, 0, 0)$ and set

$$g_\delta(x, t) = tG_\delta(x).$$

Provided that T is small enough, $F[g_\delta]$ will be close to $(0, 0)$ as required. This is the point at which we need the freedom to take our time interval $[0, T]$ to be short.

We now give more details.

Step 1 (Banach spaces of functions). Whenever U is a nonempty bounded open subset of \mathbb{R}^n or of \mathbb{R}^{n-1} and k is a positive integer and $0 < \alpha < 1$, we have Banach spaces $C^{k+\alpha}(U)$ of functions $f: U \rightarrow \mathbb{R}$ whose k th derivatives are Hölder continuous with exponent α . We denote the usual norms by $\|f\|_{C^{k+\alpha}}$. Similarly, on domains $\mathcal{M} \times [0, T]$ we have Banach spaces

$$C^{k+\alpha, (k+\alpha)/2}(U \times [0, T])$$

of functions $g: U \times [0, T] \rightarrow \mathbb{R}$ with norms denoted by $\|g\|_{C^{k+\alpha, (k+\alpha)/2}}$. These norms are those set forth by Ladyzhenskaya, Solonnikov, and Ural'ceva in [LSU, Chap. 1, § 1]; we more closely follow the terminology of Evans and Spruck [ES2, Chap. 2] (we should add the term $\langle Du \rangle_x^{(\alpha)}$ in the definition of $\|u\|_{C^{1+\alpha, (1+\alpha)/2}}$ there). In particular,

$$\begin{aligned} \|g\|_{C^{2+\alpha, (2+\alpha)/2}} = & \sup_{x,t} |g(x, t)| + \sup_{x,t} |Dg(x, t)| + \sup_{x,t} |D^2g(x, t)| + \sup_{x,t} |g_t(x, t)| \\ & + \sup \left\{ \frac{|Dg(x, s) - Dg(x, t)|}{|s - t|^{(1+\alpha)/2}} : (x, s), (x, t) \in W \times [0, T] \text{ with } s \neq t \right\} \\ & + \sup \left\{ \frac{|D^2g(x, t) - D^2g(y, t)|}{|x - y|^\alpha} : (x, t), (y, t) \in W \times [0, T] \text{ with } x \neq y \right\} \\ & + \sup \left\{ \frac{|g_t(x, t) - g_t(y, t)|}{|x - y|^\alpha} : (x, t), (y, t) \in W \times [0, T] \text{ with } x \neq y \right\} \\ & + \sup \left\{ \frac{|D^2g(x, s) - D^2g(x, t)|}{|s - t|^{\alpha/2}} : (x, s), (x, t) \in W \times [0, T] \text{ with } s \neq t \right\} \\ & + \sup \left\{ \frac{|g_t(x, s) - g_t(x, t)|}{|s - t|^{\alpha/2}} : (x, s), (x, t) \in W \times [0, T] \text{ with } s \neq t \right\}. \end{aligned}$$

As indicated above, we assume that \mathcal{M} is a compact smooth hypersurface of \mathbb{R}^n that is the boundary of a region A . By smooth, we mean that \mathcal{M} is everywhere locally the graph of a function (written in Cartesian coordinates) that is three times Hölder continuously differentiable with exponent α (some $0 < \alpha < 1$). We fix $w > 0$ sufficiently small such that, in the open set $W = \{x: |d(x)| < w\}$, the nearest point retraction mapping $\pi: W \rightarrow \mathcal{M}$ is twice Hölder continuously differentiable. Our signed distance function d is three times Hölder continuously differentiable in W .

For our purposes, on the domain \mathcal{M} , we use the Banach spaces $C^{k+\alpha}(\mathcal{M})$ of functions $f: \mathcal{M} \rightarrow \mathbb{R}$ for which $f \circ \pi \in C^{k+\alpha}(W)$ with corresponding norm given by

$$\|f\|_{C^{k+\alpha}} = \|f \circ \pi\|_{C^{k+\alpha}}.$$

Similarly, on domains $\mathcal{M} \times [0, T]$ we use the Banach spaces

$$C^{k+\alpha, (k+\alpha)/2}(\mathcal{M} \times [0, T])$$

of functions $g : \mathcal{M} \times [0, T] \rightarrow \mathbb{R}$ for which $g \circ \pi \times \mathbf{1} \in C^{k+\alpha, (k+\alpha)/2}(W \times [0, T])$ with corresponding norms given by

$$\|g\|_{C^{k+\alpha, (k+\alpha)/2}} = \|g \circ \pi \times \mathbf{1}\|_{C^{k+\alpha, (k+\alpha)/2}}.$$

Step 2 (the parabolic PDE). We assert that the nonlinear parabolic PDE that $g(x, t)$ must satisfy to correspond to a smooth Φ curvature flow is the following:

$$(7.4) \quad \begin{aligned} \frac{g_t}{|\mathbf{n} - (I - gD^2d)^{-1}\nabla g|} &= \text{trace} ([D^2\Phi(\mathbf{n} - (I - gD^2d)^{-1}\nabla g)] \\ &\quad \circ [(I - gD^2d)^{-1} \circ D^2g \\ &\quad \circ (I - gD^2d)^{-1} - (I - gD^2d)^{-1} \circ D^2d]); \end{aligned}$$

here I denotes the $n \times n$ identity matrix, $\nabla g(x, t) =: \nabla(g(\cdot, t) \circ \pi)(x)$, $D^2g(x, t) =: D^2(g(\cdot, t) \circ \pi)(x)$. This equation is a translation to our coordinates x in \mathcal{M} of the formulas given in §2.2 based on the fact that, for each time t , the surface $\{x + g(x, t)\mathbf{n}(x) : x \in \mathcal{M}\}$ has defining equation

$$d = g(\cdot, t) \circ \pi$$

in \mathbb{R}^n . The asserted equation follows from this observation, together with the facts that

$$\nabla(g(\cdot, t) \circ \pi)(z) = (I - g(\pi z, t)D^2d(\pi z))^{-1}\nabla g(\cdot, t)(\pi z)$$

and

$$\begin{aligned} D^2(g(\cdot, t))(z) \\ = (I - g(\pi z, t)D^2d(\pi z))^{-1} \circ D^2(g(\cdot, t) \circ \pi)(\pi z) \circ (I - g(\pi z, t)D^2d(\pi z))^{-1}. \end{aligned}$$

As indicated in (7.2), we abbreviate the expression on the right-hand side of the equality in (7.4) as $G(x, g, Dg, D^2g)$ and define our functions

$$F : C^{2+\alpha, (2+\alpha)/2}(\mathcal{M} \times [0, T]) \cap \{g : \sup |g| < w\} \rightarrow C^{\alpha, \alpha/2}(\mathcal{M} \times [0, T]) \times C^{2+\alpha}(\mathcal{M})$$

by formula (7.3).

Step 3 (the Fréchet derivatives). Corresponding to fixed g , we then compute the linearized operator

$$(7.5) \quad \begin{aligned} DF[g] : C^{2+\alpha, (2+\alpha)/2}(\mathcal{M} \times [0, T]) &\rightarrow C^{\alpha, \alpha/2}(\mathcal{M} \times [0, T]) \times C^{2+\alpha}(\mathcal{M}), \\ DF[g](h) &= \left[\frac{h_t}{|\mathbf{n} - (I - gD^2d)^{-1}\nabla g|} - \frac{1}{2} \frac{g_t}{|\mathbf{n} - (I - gD^2d)^{-1}\nabla g|} (\mathbf{n} - (I - gD^2d)^{-1}\nabla g) \right. \\ &\quad \cdot [-(I - gD^2d)\nabla h - (I - gD^2d)^{-2}D^2d\nabla gh] \\ &\quad - \text{trace} ([D^2\Phi(\mathbf{n} - (I - gD^2d)^{-1}\nabla g)] \\ &\quad \quad \circ [(I - gD^2d)^{-1} \circ (D^2(h \circ \pi)) \circ (I - gD^2d)^{-1}]) \\ &\quad + \text{trace} ([D^2\Phi(\mathbf{n} - (I - gD^2d)^{-1}\nabla g)] \\ &\quad \quad \circ [(I - gD^2d)^{-2} \circ D^2d \circ D^2g \circ (I - gD^2d)^{-1}h] \\ &\quad \quad + [(I - gD^2d)^{-1} \circ D^2g \circ (I - gD^2d)^{-2}(D^2d)h] \\ &\quad \quad \left. - [(I - gD^2d)^{-2} \circ (D^2d)^2h] \right) \\ &\quad + \text{trace} ([D^3\Phi(\mathbf{n} - (I - gD^2d)^{-1}\nabla g) \\ &\quad \quad \circ [(I - gD^2d)^{-1}\nabla h + (I - gD^2d)^{-2}D^2d\nabla gh] \\ &\quad \quad \circ [(I - gD^2d)^{-1} \circ (D^2(g \circ \pi)) \circ (I - gD^2d)^{-1} \\ &\quad \quad \quad - (I - gD^2d)^{-1}D^2d]), h(\cdot, 0) \Big]. \end{aligned}$$

By inspection of (7.5), we see that the linear operator $DF[g]$ depends continuously on g .

Step 4 (a bound in local coordinates on the mapping norms of the inverses of the Fréchet derivatives). We localize the linear second-order PDEs (7.5) to regions in which for some orthonormal coordinate system \mathcal{M} can be written as the graph of a function $\varphi(x_1, x_2, \dots, x_{n-1})$ defined in some region U in \mathbb{R}^{n-1} and having Hölder continuous third derivatives. If we represent h in these local coordinates $(x_1, x_2, \dots, x_{n-1})$, then the restriction of $DF[g]$ to these coordinates has the form $h \mapsto (\mathcal{L}_U^*(h), h(\cdot, 0))$. We multiply this \mathcal{L}_U^* by

$$|\mathbf{n} - (I - gD^2d)^{-1}\nabla g|$$

(which does not vanish for g 's in our domain) to obtain a partial differential operator \mathcal{L}_U in the form

$$(7.6) \quad \mathcal{L}_U(h) = \left(\frac{\partial h}{\partial t}(x, t) - \sum_{i,j=1}^{n-1} a_{i,j}(x, t) \frac{\partial^2 h}{\partial x_i \partial x_j}(x, t) + \sum_{i=1}^{n-1} b_i(x, t) \frac{\partial h}{\partial x_i}(x, t) + c(x, t)h(x, t) \right).$$

We check that the coefficient functions $a_{i,j}(x, t)$, $b_i(x, t)$, $c(x, t)$ all belong to

$$C^{0+\alpha, (0+\alpha)/2}(U \times [0, T])$$

with uniformly bounded $\|\cdot\|_{C^{0+\alpha, (0+\alpha)/2}}$ norms, provided that our g 's, which belong to

$$C^{2+\alpha, (2+\alpha)/2}(U \times [0, T]) \cap \{g: |g(x, t)| < w\},$$

have $\|g\|_{C^{2+\alpha, (2+\alpha)/2}}$ uniformly bounded. By direct computation, we infer from the ellipticity of Φ that the $a_{i,j}(x, t)$'s satisfy the ellipticity condition

$$\sum_{i,j=1}^{n-1} a_{i,j}(x, t) \xi_i \xi_j \geq \nu |\xi|^2 \quad \text{for all } \xi \in \mathbb{R}^{n-1};$$

here the positive constant ν depends only on norm bounds for our g 's. Theorem 10.1 of [LSU, Chap. 4, p. 351] applies to linear partial differential operators of the form of \mathcal{L} in (7.6), and we infer from that theorem the existence of constants $C_{T,U}$, $B_{T,U}$ (depending only on T and our norm bounds on the coefficients, and hence only on T and the norm bounds for our g 's) such that, for each

$$h \in C^{2+\alpha, (2+\alpha)/2}(U \times [0, T]),$$

$$(7.7) \quad \|h\|_{C^{2+\alpha, (2+\alpha)/2}} \leq C_{T,U} \|\mathcal{L}(h)\|_{C^{0+\alpha, (0+\alpha)/2}} + C_{T,U} \|h(\cdot, 0)\|_{C^{2+\alpha, (2+\alpha)/2}} + B_{T,U} \|h\|_{L^\infty}.$$

Estimate (7.7) seems the best local estimate. It, however, has the unfortunate $\|h\|_{L^\infty}$ dependence.

Step 5 (a global bound on the mapping norms of the inverses of the Fréchet derivatives that depends on T). Since \mathcal{M} can be covered by finitely many local coordinates patches U , we infer that the local bounds given for the mapping norms of the $DG(g)$'s give global bounds. By analogy, we write $DF[g](h) = (\mathcal{L}^*(h), h(\cdot, 0))$ and define a global operator \mathcal{L} by requiring that its restrictions be the \mathcal{L}_U 's above. Our local estimates then guarantee, for each

$$h \in C^{2+\alpha, (2+\alpha)/2}(\mathcal{M} \times [0, T]),$$

that

$$(7.8) \quad \|h\|_{C^{2+\alpha, (2+\alpha)/2}} \leq C_T \|\mathcal{L}(h)\|_{C^{0+\alpha, (0+\alpha)/2}} + C_T \|h(\cdot, 0)\|_{C^{2+\alpha, (2+\alpha)/2}} + B_T \|h\|_{L^\infty}.$$

We claim that estimate (7.8) holds with the constant B_T set equal to zero, provided that we increase the size of C_T if necessary. To see this, we consider the function

$$h_C(x, t) = e^{-Ct}h(x, t) \quad \text{for large } C$$

and let (p_C, t_C) be a point in $\mathcal{M} \times [0, T]$ at which h_C assumes a maximum value. This point (p_C, t_C) corresponds to some point (x_C, t_C) in one of our coordinate patches U . In the coordinate patch U , we evaluate

$$\mathcal{L}_U(h_C) = -C e^{-Ct}h + e^{-Ct}\mathcal{L}_U(h).$$

Either $t_C = 0$ so that the maximum of h_C and hence of h is dominated by the sup norm of $h(\cdot, 0)$, or $t_C > 0$ and we can use the first and second derivative properties of function at their maximums to conclude that

$$0 \leq -c(x_C, t_C)h_C(x_C, t_C) - C e^{-Ct_C}h(x_C, t_C) + e^{-Ct_C}\mathcal{L}_U(h)(x_C, t_C),$$

in which case (since C was large and positive) the maximum of h_C and hence of h is dominated by $\mathcal{L}_U(h)$. We can replace h by $-h$ to obtain a corresponding estimate for the minimum of h . We conclude that

$$(7.9) \quad \|h\|_{C^{2+\alpha, (2+\alpha)/2}} \leq C_T \|\mathcal{L}(h)\|_{C^{0+\alpha, (0+\alpha)/2}} + C_T \|h(\cdot, 0)\|_{C^{2+\alpha, (2+\alpha)/2}}$$

for suitable C_T 's.

Step 6 (a global bound on the mapping norms of the inverses of the Fréchet derivatives that does not depend on T). We assert that

$$(7.10) \quad \|h\|_{C^{2+\alpha, (2+\alpha)/2}} \leq C \|\mathcal{L}(h)\|_{C^{0+\alpha, (0+\alpha)/2}} + C \|h(\cdot, 0)\|_{C^{2+\alpha, (2+\alpha)/2}},$$

provided that we take the constant C to be that corresponding to $T=1$ in (7.9). This is accomplished by the following device. To the operators \mathcal{L}_U for functions defined on domains $U \times [0, T]$ above, we associate new operators defined on $U \times [0, 1]$ by setting

$$a_{i,j}(x, t) = \begin{cases} a_{i,j}(x, t) & \text{if } 0 \leq t \leq T, \\ a_{i,j}(x, T) & \text{if } T \leq t \leq 1, \end{cases}$$

and so forth. The Hölder norms of the coefficients of the \mathcal{L}_U 's on the time interval $[0, 1]$ clearly remain bounded by the norms on $[0, T]$. Our assertion follows.

This completes our proof that the $DF[g]^{-1}$'s have bounded mapping norms on their images.

Step 7 (the operators $DF[g]$ are surjections). As mentioned above, we wish to show surjectivity by using the homotopy criterion of Theorem 5.2 of [GT]. We choose a new infinitely differentiable manifold \mathcal{N} , which is very close to \mathcal{M} , and we let ρ denote the nearest point retraction mapping of a neighborhood of \mathcal{N} (containing \mathcal{M}) onto \mathcal{N} . The restriction of ρ to \mathcal{M} gives a diffeomorphism of \mathcal{M} with \mathcal{N} , which is three times Hölder continuously differentiable. This produces an isomorphism of our various function spaces on \mathcal{M} with the corresponding spaces on \mathcal{N} . We denote by $\Delta_{\mathcal{N}}$ Laplace's differential operator on \mathcal{N} . It is well known that we can solve the usual heat equation

$$\frac{\partial g}{\partial t} - \Delta_{\mathcal{N}}g = \varphi$$

with initial condition $g(\cdot, 0) = \psi$ for all smooth φ 's and ψ 's. It is difficult, however, to find a statement of this fact useful for our purposes. However, in desperation, we can invoke Theorem 5.2 [LM, p. 30] (as used in the proof [LM, Thm. 5.3, p. 32]) to infer the existence of infinitely differentiable solutions to our equation, provided that

φ and ψ are both infinitely differentiable. The Hölder estimates on $DF[g]^{-1}$'s work equally well for the inverse of the operator $\partial/\partial t - \Delta_{\mathcal{N}}$. Since we know all infinitely differentiable functions are in the range of $\partial/\partial t - \Delta_{\mathcal{N}}$, we can use these Hölder estimates to conclude that Hölder continuous φ 's and ψ 's are also appropriately in the image. Our above diffeomorphism now translates the problem with the heat operator on \mathcal{N} to a uniformly parabolic problem on \mathcal{M} (actually just a change of coordinates). This new problem on \mathcal{M} then gives the desired surjectivity property, and we conclude that each of our $DF[g]$'s is onto.

We have now shown that the $DF[g]$'s are uniformly isomorphisms as required. The remainder of argument to prove the theorem was set forth above. \square

COROLLARY 7.2 (perturbations of smooth elliptic curvature flows). *Suppose that our surface energy function Φ is smooth and elliptic. Suppose also that $[L(t)]$ is a time-parametrized family of solids such that the $\partial L(t)$'s constitute a smooth Φ curvature flow for $0 \leq t < t_0$ (some t_0). Then, for each $0 < t_1 < t_0$ there exist the following:*

- (a) *A positive initial perturbation distance bound ρ ;*
- (b) *A positive perturbation velocity bound σ ;*
- (c) *Solids $L[r, s, t] \in \mathcal{K}$ for perturbation parameters $-\rho < r < \rho$ and $-\sigma < s < \sigma$, and times $0 \leq t \leq t_1$;*
- (d) *Associated signed distance functions $d[r, s, t]: \mathbb{R}^n \rightarrow \mathbb{R}$ for $-\rho < r < \rho$, $-\sigma < s < \sigma$, and $0 \leq t \leq t_1$, such that $d[r, s, t](x) = \pm \text{dist}(x, \partial L[r, s, t])$, where we take the minus sign for $x \in L[r, s, t]$;*

with the following properties

- (1) *For each $0 \leq t \leq t_1$, $L[0, 0, t] = L(t)$;*
- (2) *For each r and s , $L[r, s, 0] = \{x: d[0, 0, 0] < r\}$;*
- (3) *For each r, s , and t , $\partial L[r, s, t]$ is a smooth hypersurface in \mathbb{R}^n that varies smoothly in the parameters r, s , and t ;*
- (4) *For each fixed r and s the normal velocity of $L[r, s, t]$ at the point p in $\partial L[r, s, t]$ and time t equals the Φ curvature plus s times the exterior normal vector, i.e.,*

$$-\frac{\partial d[r, s, t]}{\partial t}(p) = H_{\Phi, \partial L[r, s, t]}(p) + s \mathbf{n}_{L[r, s, t]}(p);$$

- (5) *For each fixed positive s , there exists $\Delta t_0 > 0$ such that, whenever $0 < \Delta t < \Delta t_0$, $0 \leq t - \Delta t < t \leq t_1$, and $p \in \partial L[r, s, t]$ then*

$$\frac{d[r, s, t - \Delta t](p)}{\Delta t} \geq \mathbf{n}_{L[r, s, t]}(p) \cdot H_{\Phi, \partial L[r, s, t]}(p) + \frac{s}{2}.$$

A similar statement holds if s is negative, except in that case we obtain

$$\frac{d[r, s, t - \Delta t](p)}{\Delta t} \leq \mathbf{n}_{L[r, s, t]}(p) \cdot H_{\Phi, \partial L[r, s, t]}(p) + \frac{s}{2}.$$

Proof. We divide the time interval $[0, t_0]$ into short subintervals $0 < t_1 < t_2 < \cdots < t_N < t_0$ of equal length. For $t_k \leq t \leq t_{k+1}$, we represent time-varying surfaces that are close to $\partial L(t_k)$ as graphs of functions $g(x, t)$ defined for $x \in \partial L(t_k)$ and $t_k \leq t \leq t_{k+1}$; such a surface is thus of the form

$$\{x + g(x, t) \mathbf{n}_{L(t_k)}(x): x \in \partial L(t_k)\}.$$

Theorem 7.1 tells us that over each suitably short time interval $[t_k, t_{k+1}]$ we are

guaranteed the existence of slight perturbations of $\partial L(t_k)$ evolving by slight perturbations of smooth Φ curvature flow. We also infer from Theorem 7.1 and the fact that $\partial L(t)$'s are evolving by smooth Φ curvature flow that surfaces that start close to $\partial L(t_k)$ at time t_k will be close to $\partial L(t_{k+1})$ at time t_{k+1} . Since smooth Φ curvature flows are viscosity Φ curvature flows, we can use the regularity theory set forth in [W] to infer that each $\partial K(t_k)$ everywhere locally will be a graph of a $C^{3+\alpha}$ function when written in Cartesian coordinates. We are thus able to apply Theorem 7.1 repeatedly for the finite number of times required to establish the present corollary. \square

LEMMA 7.3 (inside and outside barriers for \mathcal{E} minimizers). *Suppose that our surface energy function Φ is smooth and elliptic and that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt . Suppose also that $[L_0]$ and $[L]$ belong to \mathcal{K} and that ∂L_0 and ∂L are smooth hypersurfaces. We specify the signed distance functions*

$$d(x) = \pm \text{dist}(x, \partial K_0), \quad [\text{respectively, } d_*(x) = \pm \text{dist}(x, \partial L_0)],$$

where we take the negative sign if $x \in K_0$ [respectively, if $x \in L_0$], and take the positive sign otherwise.

(1) If $K_0 \subset L_0$ and $K \subset L$ and

$$\frac{d_*(p)}{\Delta t} > \mathbf{n}_L(p) \cdot H_{\Phi, \partial L}(p)$$

for each $p \in \partial L$, then $\text{spt } \partial[K] \cap \partial L = \emptyset$.

(2) If $L_0 \subset K_0$ and $L \subset K$ and

$$\frac{d_*(p)}{\Delta t} < \mathbf{n}_L(p) \cdot H_{\Phi, \partial L}(p)$$

for each $p \in \partial L$, then $\text{spt } \partial[K] \cap \partial L = \emptyset$.

Proof. We will prove conclusion (1) of Lemma 7.3 by contradiction. Suppose then there is a point p in $\text{spt } \partial[K] \cap \partial L$. The assumption that $K \subset L$ means that L is a smooth, one-sided barrier to $\text{spt } \partial[K]$. We infer from conclusion (1) of Theorem 3.10 that p is a regular point of $\text{spt } \partial[K]$. We infer from assumption (2) of Theorem 3.8 that

$$\frac{d(p)}{\Delta t} = \mathbf{n}_K(p) \cdot H_{\Phi, \partial K}(p).$$

The inclusion $K_0 \subset L_0$ implies that $d_*(p) \leq d(p)$. The inclusion $K \subset L$ implies that

$$\mathbf{n}_K(p) \cdot H_{\Phi, \partial K}(p) \leq \mathbf{n}_L(p) \cdot H_{\Phi, \partial L}(p)$$

and that $\mathbf{n}_K(p) = \mathbf{n}_L(p)$. We combine these inequalities to obtain

$$\mathbf{n}_L(p) \cdot H_{\Phi, \partial L}(p) < \frac{d_*(p)}{\Delta t} \leq \frac{d(p)}{\Delta t} = \mathbf{n}_K(p) \cdot H_{\Phi, \partial K}(p) \leq \mathbf{n}_L(p) \cdot H_{\Phi, \partial L}(p),$$

which is false. This proves (1) of Lemma 7.3. Conclusion (2) of the lemma follows with a similar argument. \square

THEOREM 7.4 (smooth elliptic curvature flows are flat flows). *Suppose that our surface energy function Φ is smooth and elliptic. Suppose also that $[L(t)]$ is a time-parametrized family of solids such that the $\partial L(t)$'s constitute a smooth Φ curvature flow for $0 \leq t < t_0$ (some t_0). Let $\partial[K(t)]$ be any flat Φ curvature flow for which $K(0) = L(0)$. Then $K(t) = L(t)$ for each $0 \leq t < t_0$.*

Proof. For $L(t)$, t_0 as above, we let t_1 , ρ , σ , $L[r, s, t]$, $d[r, s, t]$ have the meaning given in (a)–(d) of Corollary 7.2. For $j = 1, 2, 3, \dots$ and $k = 0, 1, 2, \dots$, we set $\Delta t_j = 2^{-j}$

and denote by $[K_j(k\Delta t_j)] \in \mathcal{K}$ the approximators used in constructing flat Φ curvature flows as in § 2.6, starting with $[K(0)] = [L(0)]$.

Clearly, for any s ,

$$K(0) \subset \{x: d[0, 0, 0](x) < \rho/2\} = L[\rho/2, s, 0].$$

We assert the existence of $0 < t_2 \leq t_1$ and a positive integer j_0 such that, for any $-\sigma/2 \leq s \leq \sigma/2$ and any $0 \leq t \leq t_2$,

$$(7.11) \quad \begin{aligned} K_j(k\Delta t_j) &\subset \{x: d[0, 0, 0](x) < \rho/6\} \\ &\subset \{x: d[0, 0, 0](x) < \rho/3\} \subset L[\rho/2, s, t] \end{aligned}$$

for all $j \geq j_0$, provided that k is sufficiently small so that $k\Delta t_j \leq t_2$. The left-hand inclusion follows from conclusions (1) and (3) of Theorem 5.4, while the right-hand inclusion follows from the smooth dependence of the $L[r, s, t]$'s on r, s, t .

We temporarily fix $0 < s_0 \leq \sigma/2$ and use conclusion (5) of Corollary 7.2 to choose $j_1 \geq j_0$ such that, whenever $j \geq j_1$, $-\rho/2 \leq r \leq \rho/2$, $\Delta t_j \leq t \leq t_1$, then

$$(7.12) \quad \frac{d[r, s_0, t - \Delta t_j](p)}{\Delta t_j} \geq \mathbf{n}_{L[r, s_0, t]}(p) \cdot H_{\Phi, \partial L[r, s_0, t]}(p) + \frac{s_0}{2}.$$

We now fix $j_2 \geq j_1$ and let k_0 denote the largest integer for which $k_0\Delta t_{j_2} \leq t_2$. By our choice (7.11), above, each $K_{j_2}(k\Delta t_{j_2})$ lies strictly inside $L[\rho/2, s_0, k\Delta t_{j_2}]$ for $k = 0, \dots, k_0$. With s_0, j_2 fixed, and k restricted to $\{0, \dots, k_0\}$, we denote by $0 \leq r_1 < \rho$ the smallest nonnegative number such that $K_{j_2}(k\Delta t_{j_2})$ lies strictly inside $L[r, s_0, k\Delta t_{j_2}]$ for each $k \in \{0, \dots, k_0\}$ and all $r_1 < r \leq \rho/2$. We assert that $r_1 = 0$. If this were not the case, there would be $k_1 \in \{1, \dots, k_0\}$ (clearly, $k_1 \neq 0$) such that $[K_{j_2}(k_1\Delta t_{j_2})]$ is an \mathcal{E} minimizer for $[K_{j_2}((k_1 - 1)\Delta t_{j_2})]$ over Δt_{j_2} with

$$K_{j_2}((k_1 - 1)\Delta t_{j_2}) \subset L[r_1, s_0, (k_1 - 1)\Delta t_{j_2}], \quad K_{j_2}(k_1\Delta t_{j_2}) \subset L[r_1, s_0, k_1\Delta t_{j_2}]$$

and

$$\text{spt } \partial[K_{j_2}(k_1\Delta t_{j_2})] \cap \partial L[r_1, s_0, k_1\Delta t_{j_2}] \neq \emptyset$$

(since otherwise there would be a smaller r_1). In view of (7.12), this situation is incompatible with conclusion (1) of Lemma 7.3. We therefore infer that $r_1 = 0$. We further infer, for each fixed $0 < s \leq \sigma/2$, that $K_j(k\Delta t_j) \subset L[0, s, k\Delta t_j]$ for all sufficiently large j and all k for which $0 \leq k\Delta t_j \leq t_2$.

An analogous argument based on conclusion (2) of Lemma 7.3 implies that, for each fixed $-\sigma/2 < s < 0$, $K_j(k\Delta t_j) \supset L[0, s, k\Delta t_j]$ for all sufficiently large j and all k for which $0 \leq k\Delta t_j \leq t_2$.

We then infer from the smooth dependence of $L[r, s, t]$ on r, s, t and the definition of flat convergence that $K(t) = L[0, 0, t] = L(t)$ for each $0 \leq t \leq t_2$.

We now wish to remedy the difficulty that $K(t) = L(t)$ only for $0 \leq t \leq t_2$. This difficulty came from the requirement that $K_j(k\Delta t_j) \subset L[\rho/2, s_0, k\Delta t_j]$ for our times Δt_j , and initially we could infer such inclusion only for $k\Delta t_j$'s not exceeding our t_2 . We now know that $\partial K_j(k\Delta t_j)$ lies extremely close to $\partial L(t_2)$ for $k\Delta t_j$'s extremely close to t_2 and j extremely large. We can now use this fact, together with shrinking Wulff shape barrier estimates given in conclusions (1) and (3) of Theorem 5.4 to infer the existence of a time t_3 greater than t_2 such that $\partial K_j(k\Delta t_j)$ lies still quite close to $\partial L(t_2)$ for $k\Delta t_j$'s between t_2 and t_3 . For huge j 's and t_3 , which is not too much greater than t_2 , we conclude that $K_j(k\Delta t_j) \subset L[\rho/2, s_0, k\Delta t_j]$ for $k\Delta t_j$'s not exceeding our t_3 .

We then look at even larger j 's to extend the arguments to a t_4 greater than t_3 . This procedure works (since our estimates are effectively uniform out to time t_1) to push the time of coincidence out to time t_1 . Since t_1 was any number less than t_0 , we conclude that $K(t) = L(t)$ for $0 \leq t < t_0$. \square

8. Examples of flat curvature flows.

8.1. Circles moving by flat arc length flow. The unique flat arc length curvature flow of curves in the plane starting with $\partial[\mathbb{B}^2((0, 0), r_0)]$ is given by

$$\partial[K(t)](t) = \begin{cases} \partial[\mathbb{B}^2((0, 0), (r_0^2 - 2t)^{1/2})] & \text{for } 0 \leq t < \frac{r_0^2}{2} \\ 0 & \text{for } \frac{r_0^2}{2} \leq t < \infty. \end{cases}$$

To see this, we suppose that $n = 2$ so that the integrand $\Phi(v) = |v|$ is the arc length integrand. Using the terminology of § 5.1 and Theorem 5.4, that we check that $W_1 = \mathbb{B}^2((0, 0), \pi^{-1/2})$, $w_* = w^* = \pi^{-1/2}$, $\Phi(\partial[W_1]) = 2\pi^{1/2}$, $c_0 = c^0 = 2\pi$, $c_1 = (32\pi)^{-1}$, $S_0(t) = S^0(t) = (S^2 - 2\pi t)^{1/2}$, and

$$W_S = \mathbb{B}^2((0, 0), \pi^{-1/2}S) \quad \text{for each } S.$$

We infer from conclusions (2), (4), and (6) of Theorem 5.4 that, for each point p and each initial radius r_0 , the shrinking circles

$$\partial\mathbb{B}^2(p, (r_0^2 - 2t)^{1/2}) \quad \text{for } 0 \leq t < \frac{r_0^2}{2}$$

are barriers to any flat arc length curvature flow from both the outside and inside. Our original assertion above follows.

8.2. Nonuniqueness of flat arc length curvature flows for an initial curve resembling a figure-eight curve. We will construct an initial solid $[K(0)]$ whose boundary curve $C = \partial K(0)$ is a continuously differentiable immersion of a circle resembling a figure eight lying on its side. The crossing occurs at the origin and has tangent given by the lines $y = \pm x$. Corresponding to one subsequence of approximating flows, the crossing in the limit flat flow will split horizontally into two halves. Corresponding to another subsequence, the crossing will split vertically. The difference in splitting occurs because on the scale appropriate to the first subsequence of Δt 's the crossing seems to occur at an angle slightly less than 90° (which favors the horizontal split for approximators associated with such Δt 's), while, on the scale appropriate to the second subsequence of Δt 's, the crossing angle is slightly greater than 90° (favoring the vertical split). The construction and argument require several steps. To be specific, we assume that $K(0)$ contains the disks $\mathbb{B}^2((12, 0), 3)$ and $\mathbb{B}^2((-12, 0), 3)$ and is disjoint from the disks $\mathbb{B}^2((0, 12), 3)$ and $\mathbb{B}^2((0, -12), 3)$.

We denote by $\xi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ reflection across the x axis and denote by $\eta: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ reflection across the y axis in \mathbb{R}^2 .

PROPOSITION 8.2.1 (approximators can be chosen with ξ and η symmetry). *Suppose that $[K_0]$ is a solid in \mathcal{K} with $K_0 = \xi K_0 = \eta K_0$, and $0 < \Delta t < \infty$. Then at least one of the currents $[K]$, which is an \mathcal{E} minimizer for $[K_0]$ over Δt , has the symmetries $K = \xi K = \eta K$.*

Proof. For any minimizing, K_* , select the quadrant Q in which

$$\Phi(\partial K_* \llcorner Q) + \frac{1}{\Delta t} \int_{x \in Q \cap K_* \Delta K_0} \text{dist}(x, \partial K_0) d\mathcal{L}^2 x$$

is the smallest and replace K_* by $K = K_* \cup \xi(K_*) \cup \eta(K_* \cup \xi(K_*))$. The main observation in showing

$$\mathcal{E}([K_0], [K], \Delta t) \leq \mathcal{E}([K_0], [K_*], \Delta t)$$

is that, for each point p in \mathbb{R}^2 , some closest point in ∂K_0 to p lies in the quadrant to which p belongs. \square

We denote by $[K_j(k\Delta t_j)]$, $j = 1, 2, 3, \dots$, $k = 0, 1, 2, \dots$, approximators to flat arc length curvature flows $[K(t)]$ beginning with $[K(0)]$; here $\Delta t_j = 2^{-j}$ for each j . In view of Proposition 8.2.1, we require that $K_j(k\Delta t_j) = \xi K_j(k\Delta t_j) = \eta K_j(k\Delta t_j)$ for each j and k . We recall the scaling estimates set forth in conclusion (1) of Theorem 3.9 and define $K_j^*(k) = \Delta t_j^{-1/2} K_j(k\Delta t)$ for each j and k (homothetic expansion by factor $\Delta t_j^{-1/2}$); hence each $[K_j^*(k+1)]$ is an \mathcal{E} minimizer for $[K_j^*(k)]$ over 1 (i.e., the time step equals 1).

8.2.2. Dumbbell-shaped barriers. We want to consider regions consisting of two widely separated large disks joined by a narrow strip. With this in mind we set

$$A(l, r, s) = \mathbb{B}^2((0, l), r) \cup \mathbb{B}^2((0, -l), r) \cup [-s, s] \times [-l, l]$$

for numbers $l > r \gg s$. In our applications, we will take $l = 12$ and $1 \leq r \leq 3$ or $\Delta t_j^{-1/2}$ times these numbers.

PROPOSITION 8.2.3 (nonshrinking central strips in the dumbbell barriers). *Suppose that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt and that $l > r \gg s > 32\pi\Delta t^{1/2}$. If $A(l, r, s) \cap K_0 = \emptyset$, then $A(l, r - (4\pi/r)\Delta t, s) \cap K = \emptyset$.*

Proof. It follows from conclusion (3) of Theorem 5.3 that $A(l, r - (4\pi/r)\Delta t, s - (4\pi/s)\Delta t) \cap K_0 = \emptyset$. Verification of the larger avoidance is based on construction of comparison curves in the \mathcal{E} minimization; details are left to the reader. \square

Analogous statements hold for the image of $A(l, r, s)$ and K_0 under any Euclidean motion or if $A(l, r, s) \subset K_0$.

The presence of two large disks at the ends of the strip means that the strip does not have to shrink to remain a barrier and that the separation persists until the disks have shrunk.

Proposition 8.2.3, together with conclusion (2) of Theorem 5.4, imply that, if Δt_j is small and

$$K_j(k_0\Delta t_j) \cap A(12, 2, 33\pi\Delta t_j^{1/2}) = \emptyset,$$

then

$$K_j((k_0 + k)\Delta t_j) \cap A(12, 1, 33\pi\Delta t_j^{1/2}) = \emptyset,$$

provided that $2\pi k\Delta t_j \leq 1$. In other words, if, for some j , $K_j(k_0\Delta t_j)$ achieves a horizontal separation by amount $33\pi\Delta t_j^{1/2}$ and $k_0\Delta t_j$ is small enough so that $K_j(k_0\Delta t_j)$ is still disjoint from the disks $\mathbb{B}^2((0, 12), 2)$ and $\mathbb{B}^2((0, -12), 2)$, then the horizontal separation in the $K_j(k\Delta t_j)$'s will persist for additional time at least $1/2\pi$.

If the separation criterion is stated in terms of the sets $K_j^*(k)$, then horizontal separation of 33π is sufficient for the separation to persist for time $1/2\pi$.

8.2.4. Double-sector-shaped regions and retraction functions. We want to consider regions shaped like two symmetric pieces of pie having vertex angles slightly smaller or slightly larger than a right angle and also retractions of the plane less the y axis onto such regions. For this section only, we use polar coordinates r, θ for the plane. For a (large) radius R and a (small) perturbation angle $-\pi/8 < \omega < \pi/8$, we set

$$X(R, \omega) = \mathbb{B}^2((0, 0), R) \cap \{(r, \theta): \text{either } |\theta| \leq \pi/4 - \omega \text{ or } |\theta - \pi| \leq \pi/4 - \omega\}.$$

We define functions $F[\omega]: \mathbb{R}^2 \sim (y \text{ axis}) \rightarrow \mathbb{R}^2$, which retract each R disk (with the y axis removed) onto $X(R, \omega)$, by requiring for each r and θ that

$$F[\omega](r, \theta) = \begin{cases} (r, \theta) & \text{if } (r, \theta) \in X(R, \omega), \\ \left(r, \left[\frac{\pi}{4} - \omega \right] - \frac{\frac{\pi}{4} - \omega}{\frac{\pi}{4} + \omega} \left[\theta - \left(\frac{\pi}{4} - \omega \right) \right] \right) & \text{if } \frac{\pi}{4} - \omega < \theta < \frac{\pi}{2}, \\ \left(r, \left[\frac{-\pi}{4} + \omega \right] + \frac{\frac{\pi}{4} - \omega}{\frac{\pi}{4} + \omega} \left[\left(-\frac{\pi}{4} + \omega \right) - \theta \right] \right) & \text{if } \frac{-\pi}{2} < \theta < -\frac{\pi}{4} + \omega, \\ \left(r, \left[\frac{3\pi}{4} + \omega \right] + \frac{\frac{\pi}{4} - \omega}{\frac{\pi}{4} + \omega} \left[\left(\frac{3\pi}{4} + \omega \right) - \theta \right] \right) & \text{if } \frac{\pi}{2} < \theta < \frac{3\pi}{4} + \omega, \\ \left(r, \left[\frac{-3\pi}{4} - \omega \right] - \frac{\frac{\pi}{4} - \omega}{\frac{\pi}{4} + \omega} \left[\theta - \left(\frac{3\pi}{4} - \omega \right) \right] \right) & \text{if } \frac{-3\pi}{4} - \omega < \theta < -\frac{\pi}{2}. \end{cases}$$

When restricted to each of the half-spaces $-\pi < \theta < \pi$ and $\pi < \theta < 3\pi/2$, our mapping $F[\omega]$ clearly is length nonincreasing and is area nonincreasing in case $\omega > 0$.

We can use the $X(R, \omega)$'s and $F[\omega]$'s to obtain estimates necessary for our construction.

We denote by k_0 the smallest integer for which $(k_0 - 1)(4/67) > 33\pi$ for use in the following proposition and additionally below.

PROPOSITION 8.2.5 (small separations grow to big separations). *Suppose that $0 < \omega < \pi/8$ and*

$$K_j^*(k) \cap \mathbb{B}^2((0, 0), 100\pi) \subset X(100\pi, \omega) \sim \mathbb{B}^2((0, 0), R)$$

for some $0 < R < 1$ and for $k = 1, 2, \dots, k_0$. Then

$$\begin{aligned} K_j^*(k) \cap \mathbb{B}^2((0, 0), 67\pi) \\ \subset \mathbb{B}^2((67\pi, 0), 67\pi - \frac{4}{67}(k-1)) \cup \mathbb{B}^2((-67\pi, 0), 67\pi - \frac{4}{67}(k-1)) \end{aligned}$$

for each $k = 0, 1, \dots, k_0$. In particular,

$$K_j^*(k_0) \cap \mathbb{B}^2((0, 0), 100\pi) \cap [-33\pi, 33\pi] \times \mathbb{R} = \emptyset.$$

Proof. The comparison surface construction used in the proof of part (5) of Theorem 5.4 is local and applies to the present case.

Analogous statements hold with $\omega < 0$, provided that K_0 is replaced by $\mathbb{R}^2 \sim K_0$, K is replaced by $\mathbb{R}^2 \sim K$, $X(S + \Delta S, \omega)$ is replaced by $X(S + \Delta S, \omega) \sim K_0$, and so forth.

PROPOSITION 8.2.6 (a sufficient condition to keep \mathcal{E} minimizing K 's inside X 's). *Suppose that $0 < \omega < \pi/8$ and $0 < S < \infty$. Suppose also that $[K]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt with $K = \xi K = \eta K$. In case*

$$K_0 \cap \mathbb{B}^2((0, 0), 2S) \subset X(2S, \omega)$$

and

$$K \cap [\mathbb{B}^2((0, 0), 2S) \sim \mathbb{B}^2((0, 0), S)] \subset X(2S, \omega),$$

then $K \cap \mathbb{B}^2((0, 0), S) \subset X(S, \omega)$.

Proof. Suppose that $[K_*]$ is an \mathcal{E} minimizer for $[K_0]$ over Δt with $K_* = \xi K_* = \eta K_*$ and with $K_* \sim X(S + \Delta S, \omega) \pm \emptyset$. Consider as a comparison solid the current associated with a set $K_* \sim \mathbb{B}^2((0, 0), S) \cup N_1$; here N_1 is obtained as follows. Write

$$F[\omega]_*[K \cap \{(x, y): x > 0\}] + F[\omega]_*[K \cap \{(x, y): x < 0\}] = \sum_i [N_i] - \sum_j [L_j]$$

in accordance with § 3.1.4 and take that N_1 . We check that

$$\mathcal{E}([K_0], [K_1], \Delta t) < \mathcal{E}([K_0], [K_*], \Delta t),$$

contrary to assumption. \square

Analogous statements hold with $\omega > 0$, provided that K_0 is replaced by $\mathbb{R}^2 \sim K_0$, K is replaced by $\mathbb{R}^2 \sim K$, $X(S + \Delta S, \omega)$ is replaced by $X(S + \Delta S, \omega) \sim K_0$, and so forth.

PROPOSITION 8.2.7 (how to satisfy the hypotheses of Proposition 8.2.6). *Suppose that $0 < \omega < \pi/8$ is given and S is chosen sufficiently large so that*

$$S \geq 100\pi \quad \text{and} \quad \frac{4\pi}{S/2} (k_0 + 1) \leq \frac{\omega S}{2}.$$

Then there exists $0 < R < S$ with the following property. Suppose, for a particular choice of j , that we know that $K_j^(0) \cap [\mathbb{B}^2((0, 0), 5S) \sim \mathbb{B}^2((0, 0), R)] \subset X(5S, 2\omega)$. Then $K_j^*(k) \cap [\mathbb{B}^2((0, 0), 2S) \sim \mathbb{B}^2((0, 0), S)] \subset X(2S, \omega)$ for all $k = 0, 1, \dots, k_0$.*

Proof. For points p lying in $\partial X(5S, 2\omega) \cap [\mathbb{B}^2((0, 0), 2S) \sim \mathbb{B}^2((0, 0), S)]$, denote by $r(p)$ the radius for which the point $q(p) = r(p)\mathbf{n}_{X(5S, 2\omega)}$ lies in the y axis. Clearly, $S/2 < |r(p)| < 3S$. Since the balls $\mathbb{B}^2(q(p), r(p))$ now lie outside $X(5S, 2\omega) \cup \mathbb{B}^2((0, 0), R)$ (provided R is small), we can use conclusion (1) of Theorem 5.4 to let the balls shrink while remaining barriers to the $K_j^*(k)$'s. Our numbers above have been chosen so that the radii $r(p)$ need shrink by amount no more than $S\omega/2$. \square

PROPOSITION 8.2.8 (initial separation). *Suppose that $0 < \omega < \pi/8$ is given and S is the smallest number for which*

$$S \geq 100\pi \quad \text{and} \quad \frac{4\pi}{S/2} (k_0 + 1) \leq \frac{\omega S}{2}.$$

Then there exists $0 < R < S$ with the following property. Suppose that K_R is an \mathcal{E} minimizer for K_0 over 1 with $K_0 = \xi K_0 = \eta K_0$, $K_R = \xi K_R = \eta K_R$, and

$$K_0 \cap [B^2((0, 0), 5S) \sim \mathbb{B}^2((0, 0), R)] = X(5S, 2\omega) \sim \mathbb{B}^2((0, 0), R).$$

Then

$$K_R \cap \mathbb{B}^2((0, 0), S) \subset X(S, \omega).$$

Proof. Corresponding to $R = 0$, we use Propositions 8.2.6 and 8.2.7, together with part (5) of Theorem 5.4 (as in the proof of Proposition 8.2.5), to infer that, if K_* is an \mathcal{E} minimizer for K_0 over 1 with $K_* = \xi K_* = \eta K_*$ and $K_0 \cap B^2((0, 0), 5S) = X(5S, 2\omega)$, then $K_* \cap \mathbb{B}^2((0, 0), S) \subset X(S, \omega)$ and

$$K_* \cap \mathbb{B}^2((0, 0), 67\pi) \subset \mathbb{B}^2((67\pi, 0), 67\pi - 4/67) \cup \mathbb{B}^2((-67\pi, 0), 67\pi - 4/67).$$

Any sequence of R 's converging to zero must contain a subsequence such the $[K_R]$'s converge to a minimizing $[K_*]$. From the uniform lower density bound, we conclude the convergence of the ∂K_R 's to ∂K_* in Hausdorff distance. The remainder of the proof is left to the reader. \square

8.2.9. An example of nonuniqueness of flat Φ curvature flows. As indicated above, we will describe the construction a special initial solid $[K]$ for which there is more than one flat arc length curvature flow beginning with $\partial[K]$. The boundary curve $C = \partial K$ will resemble a figure-eight curve C lying on its side and passing through the origin tangent to the lines $y = \pm x$. C will also have the symmetries $C = \xi C = \eta C$. First, we specify small perturbation angles $\omega_i = (-1)^i 2^{-(100+i)}$ of alternating signs. They key to the construction of C near the origin is selection of special radii $S_1 > R_1 > S_2 > R_2 \dots$

converging to zero; the proposition below asserts that we can make a suitable selection. Having suitably selected such radii, we construct a symmetric C such that, for each i ,

$$\begin{aligned} C \cap [\mathbb{B}^2((0, 0), S_i) \sim \mathbb{B}^2((0, 0), R_i)] \\ = \partial X(S_i + 1, 2\omega_i) \cap [\mathbb{B}^2((0, 0), S_i) \sim \mathbb{B}^2((0, 0), R_i)], \end{aligned}$$

while

$$C \cap [\mathbb{B}^2((0, 0), R_i) \sim \mathbb{B}^2((0, 0), S_{i+1})]$$

is a reasonable interpolation curve. We then extend C outside $\mathbb{B}^2((0, 0), S_1)$ to be a smooth symmetric curve resembling a figure eight on its side so that, if $K(0)$ be the bounded region whose topological boundary is C , then $K(0)$ contains the disks $\mathbb{B}^2((12, 0), 3)$ and $\mathbb{B}^2((-12, 0), 3)$ and is disjoint from the disks $\mathbb{B}^2((0, 12), 3)$ and $\mathbb{B}^2((0, -12), 3)$.

Corresponding to a suitable choice of $j(1) < j(2) < j(3) < \cdots$, we let $\partial[K_{j(i)}(k\Delta t_{j(i)})]$ be any collection of approximate flat Φ curvature flows for which

$$K(k\Delta t_{j(i)}) = \xi K(k\Delta t_{j(i)}) = \eta K(k\Delta t_{j(i)}).$$

We then have the following proposition.

PROPOSITION 8.2.10 (nonuniqueness of flat arc length curvature flows). *It is possible to choose $j(i)$'s, S_i 's, and R_i 's with which to make the construction in § 8.2.9 so that the following conditions hold.*

(1) *For all sufficiently large i 's that are even, our approximating boundary curves $\partial[K_{j(i)}(k\Delta t_{j(i)})]$ can be chosen to split horizontally into two pieces, one in the right half plane and the other in the left half plane, and this type of horizontal split of the crossing will persist until the final disappearance of the evolving curves. Furthermore, if $\partial[K(t)]$ is a flat arc length curvature flow that is the limit of a subsequence of $K_{j(i)}(\cdot)$'s in which the i 's are even, and $C(t) = \partial K(t)$, then $C(t)$ instantly splits horizontally into at least two components, which remain separated until the final disappearance.*

(2) *For all sufficiently large i 's that are odd, our approximating boundary curves $\partial[K_{j(i)}(k\Delta t_{j(i)})]$ can be chosen to split vertically, and near the origin there will be part in the upper half plane and part in the lower half plane. This type of vertical split of the crossing will remain at least while $k\Delta t_{j(i)} \leq 1/2\pi$. Furthermore, if $\partial[K(t)]$ is a flat arc length curvature flow that is the limit of a subsequence of $K_{j(i)}(\cdot)$'s in which the i 's are odd, and $C(t) = \partial K(t)$, then $C(t)$ instantly splits vertically and, in small neighborhoods of the origin, $C(t)$ contain at least two components separated vertically. Vertical separation near the origin persists for time at least $1/2\pi$.*

Proof. We pick $j(1)$ and $0 < S_1 < 1$ such that, if $5S = \Delta t_{j(1)}^{-1/2} S_1$, then S is the smallest number for which

$$S \geq 100\pi \quad \text{and} \quad \frac{4\pi}{S/2} (k_0 + 1) \leq \frac{\omega_1 S}{2}.$$

Associated with this S and with $\omega = \omega_1$, we choose $R_1 < S_1$ so that the number $\Delta t_{j(1)}^{-1/2} R_1$ satisfies the requirements of the number R in Propositions 8.2.7 and 8.2.8. We then pick $j(2) \gg j(1)$ and $S_2 \gg R_1$ to satisfy the analogous conditions above. We then sequentially pick R_2, S_3, R_3 , and so forth. We use Propositions 8.2.5–8.2.8 to establish the asserted behavior for time at least $1/2\pi$. In the case of the horizontal split, we can additionally use large circle barriers to guarantee that rejoining does not occur. \square

8.3. How we might construct an initial surface for which there are a continuum of different flat surface area curvature flows. We take $n = 3$ so that $\Phi(v) = |v|$ is the surface

area integrand. We suggest how to construct an initial surface S for which there might be a continuum of different flat curvature flows. The basic idea is to take a figure-eight curve C as in § 8.2, translate it a substantial x distance away from the origin, and then rotate the image around the y axis to obtain a surface S of revolution. It seems that careful modification of S near its singular curve (inspired by the construction of § 8.2.9 and Proposition 8.2.10) would lead to flat surface area curvature flows, starting with S , in which there was initial horizontal separation for positive x and initial vertical separation for negative x . In particular, the type of separation changes from horizontal to vertical, or vice versa, as we cross the yz plane along the singular curve. This being the case, we apparently could further modify the behavior of S near its singular curve so that the plane at which type of separation changes could be any plane containing the z axis, provided that we used the right subsequence of approximations. Much more elaborate types of splitting patterns also seem to be possible.

Acknowledgments. The authors thank Andrew Roosen and Andrea Sufke for a careful reading of a large part of this manuscript and for their suggestions for the improvement of the exposition.

REFERENCES

- [A1] F. ALMGREN, *Existence and regularity almost everywhere of solutions to elliptic variational problems with constraints*, Mem. Amer. Math. Soc., 165 (1976), vii, 199.
- [A2] ———, *Deformations and multiple-valued functions*, in Geometric Measure Theory and the Calculus of Variations, Proc. Sympos. Pure Math., 44 (1986), pp. 29–130.
- [A3] ———, *Optimal isoperimetric inequalities*, Indiana Univ. Math. J., 35 (1986), pp. 451–547.
- [Al] R. ALMGREN, *Variational algorithms and pattern formation in dendritic solidification*, J. Comp. Physics (1993), to appear.
- [ASS] F. ALMGREN, R. SCHOEN, AND L. SIMON, *Regularity and singularity estimates for hypersurfaces minimizing parametric elliptic variational integrals*, Acta Math., 139 1(1977), 527–538.
- [AT] F. ALMGREN AND J. E. TAYLOR, *Flat flow is motion by crystalline curvature for curves with crystalline energies*, preprint, The Geometry Center, 1992.
- [AW] F. ALMGREN AND L. WANG, *Mathematical existence of crystal growth with Gibbs–Thomson curvature effects*, in preparation.
- [B] M. S. BERGER, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.
- [Bo] E. Bombieri, *Regularity theory for almost minimal currents*, Arch. Rational Mech. Anal., 78 (1982), pp. 99–130.
- [BSS] G. BARLES, H. M. SONER, AND P. SOUGANIDIS, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), this issue, pp. 439–469.
- [CGG] Y.-G. CHEN, Y. GIGA, AND S. GOTO, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geometry, 33 (1991), pp. 749–786.
- [CW] L. CAFFARELLI AND L. WANG, *Harnack inequality approach to the interior regularity of parabolic equations*, Indiana Univ. Math. J., to appear.
- [DS] G. DZIUK AND A. SCHMIDT, personal communication.
- [ES1] L. C. EVANS AND J. SPRUCK, *Motion of level sets by mean curvature I*, J. Differential Geometry, 33 (1991), pp. 635–681.
- [ES2] ———, *Motion of level sets by mean curvature II*, Trans. Amer. Math. Soc., to appear.
- [F] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, Berlin, New York, 1969.
- [FF] H. FERDER AND W. H. FLEMING, *Normal and integral currents*, Ann. Math., 72 (1960), pp. 458–520.
- [G] M. GIAQUINTA, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*, Princeton Univ. Press, Princeton, NJ, 1983.
- [GG] Y. GIGA AND S. GOTO, *Geometric evolution of phase-boundaries*, Mathematics and Its Applications 43, Springer-Verlag, Berlin, New York, 1992.
- [GT] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.
- [Gu] M. GURTIN, *Thermomechanics of Evolving Phase Boundaries*, preprint.
- [KP] S. KRANTZ AND H. PARKS, *Distance to C^k hypersurfaces*, J. Differential Equations, 40 (1981), pp. 116–120.

- [LS] S. LUCKHAUS AND T. STURZENHECKER, *An implicit time discretization for mean curvature flow*, preprint.
- [LSU] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, in *Translations of Mathematical Monographs*, Vol. 23, Amer. Math. Soc., Providence, RI, 1968.
- [LU] O. A. LADYZHENSKAYA AND N. N. URAL'CEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
- [LM] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications II*, Springer-Verlag, New York, 1972.
- [M1] F. MORGAN, *Geometric Measure Theory. A Beginner's Guide*, Academic Press, New York, 1987.
- [M2] ———, *The cone over the Clifford torus in \mathbb{R}^4 is Φ -minimizing*, *Math. Ann.*, 289 (1991), pp. 341–354.
- [R] A. ROOSEN, *Simulation of two-dimensional faceted crystal growth in a single diffusion field*, Computational Crystal Growers Workshop, Selected Lectures in Mathematics, Amer. Math. Soc., Providence, RI, 1992, pp. 89–91; video, 00:14:10–00:20:25.
- [RT] A. ROOSEN AND J. E. TAYLOR, *Simulation of crystal growth with faceted interfaces*, *Interface Dynamics and Growth*, *Mat. Res. Soc. Symp. Proc.*, 237 (1992), pp. 25–36.
- [S] L. SIMON, *Lectures on Geometric Measure Theory*, *Proc. of the Centre for Mathematical Analysis*, Vol. 3, Australian National University, 1983.
- [T1] J. E. TAYLOR, *Unique structure of solutions to a class of nonelliptic variational problems*, *Proc. Sympos. Pure Math.*, 27 (1974), pp. 481–489.
- [T2] ———, *Motion of curves by crystalline curvature including triple junctions and boundary points*, in *Differential Geometry*, *Proc. Sympos. Pure Math.*, to appear.
- [T3] ———, *Motion by crystalline curvature*, in *Computing Optimal Geometries*, *Selected Lectures in Mathematics*, Amer. Math. Soc., Providence, RI, 1991, pp. 63–65. (Including video.)
- [T4] ———, *Geometric crystal growth in 3D via faceted interfaces*, *Computational Crystal Grower Workshop*, *Selected Lectures in Mathematics*, Amer. Math. Soc., Providence, RI, 1992, pp. 111–113; video, 00:20:25–00:26:00.
- [TCH] J. E. TAYLOR, J. W. CAHN AND C. A. HANDWERKER, *Geometric models of crystal growth*, *Acta metall. mater.*, 40 (1992), pp. 1443–1474.
- [U] A. UNDERWOOD, *Polyhedral Mean Curvature and Its Relationship to Smooth Mean Curvature*, informal notes, Princeton University, Princeton, NJ, 1992.
- [W] L. WANG, *On the regularity theory of fully nonlinear parabolic equations: II*, *Comm. Pure Appl. Math.*, 45 (1992), pp. 141–178.

FRONT PROPAGATION AND PHASE FIELD THEORY*

G. BARLES[†], H. M. SONER[‡], AND P. E. SOUGANIDIS[§]

This paper is dedicated to Wendell Fleming on the occasion of his 65th birthday.

Abstract. The connection between the weak theories for a class of geometric equations and the asymptotics of appropriately rescaled reaction-diffusion equations is rigorously established. Two different scalings are studied. In the first, the limiting geometric equation is a first-order equation; in the second, it is a generalization of the mean curvature equation. Intrinsic definitions for the geometric equations are obtained, and uniqueness under a geometric condition on the initial surface is proved. In particular, in the case of the mean curvature equation, this condition is satisfied by surfaces that are strictly starshaped, that have positive mean curvature, or that satisfy a condition that interpolates between the positive mean curvature and the starshape conditions.

Key words. viscosity solutions, mean curvature flow, front propagation, reaction-diffusion equations

AMS(MOS) subject classifications. 35A05, 35K57, 53A10

Introduction. In this paper we study the connection between the *weak propagation of fronts* (closed hypersurfaces in \mathbb{R}^N , which propagate in the normal direction with the velocity depending on the position, the normal vector, and its gradient) and the *phase field theory*, as it applies to the study of the asymptotic behavior of reaction-diffusion equations. More specifically, we study the properties of the *signed distance function* to the front; we relate these properties to the level set formulation of moving fronts, and we present some new, general, and, in some cases, sharp results guaranteeing the uniqueness of the fronts (“no interior”). Finally, we develop a rigorous justification of the “phase field” theory.

The study of propagating fronts is very interesting from both the theoretical point of view as well as for applications (e.g., phase transitions in continuum mechanics, flame propagation, pattern formation, chemical kinetics, etc.). The strong geometrical formulation of the motion (which requires smoothness) faces the development of singularities; the motion can, therefore, be defined only locally in time, which is quite unsatisfactory for the applications. On the other hand, a weak geometrical formulation by Brakke [Br] for motion by mean curvature gave rise to nonuniqueness problems, but resulted in deep regularity results for the motion. More recently, two different approaches were introduced to deal with these issues, namely, the level set and the phase field approach. The level set approach, which was put forward by Evans and Spruck [ESp1] for motion by mean curvature and Chen, Giga, and Goto [CGG] for general motions, is based on considering the front as a level set (for definiteness the zero level set) of the solution of a degenerate parabolic partial differential equation (pde). The phase field approach, suggested by Bronsard and Kohn [BrK] and DeGiorgi

* Received by the editors February 28, 1992; accepted for publication (in revised form) July 1, 1992.

[†] Faculté des Sciences de Techniques, Université de Tours, Parc de Grandmont, 37200 Tours, France. The work of this author was partially supported by PICS 955702.

[‡] Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. The work of this author was partially supported by National Science Foundation grant DMS-9002249 and by the Army Research Office through the Center for Nonlinear Analysis.

[§] Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was partially supported by National Science Foundation grants DMS-8801208, DMS-9024617, and DMS-8657464 (PYI), Army Research Office contract DAAL03-90-G-0012, and the Sloan Foundation. Part of this work was done while the author was visiting the University of Paris-Dauphine.

[D], defines the front as the boundary of the regions where the solutions of certain (scaled) reaction diffusion equations converge to the equilibria points of the associated vector field. Both approaches have their own advantages. The level set formulation provides a large number of analytical tools to study the motion because it allows for the use of very recent developments of the theory of nonlinear degenerate parabolic pde's. The phase field formulation is very indirect but also closely related to (and very natural for) the applications. A great deal of work in this paper is devoted to justifying the "phase field" formulation. One way to relate these two approaches is to study the properties of the distance function to the front; much of the work in this paper is devoted to this. In fact, we could propose an alternative way to study front propagation using the distance function. This was done by Soner [So] when the normal velocity of the front does not depend on its position. We chose not to do so in this paper, although given what we prove here for the distance function we can easily develop such an approach. A very intriguing mathematical question arising with the weak formulation of moving fronts is whether such fronts are uniquely determined by their initial position (if they are described using the distance function); this is closely related to whether the level set formulation gives rise to *fat* level sets. Two sections in this paper are devoted to studying these questions.

The paper is organized as follows: In § 1 we recall the level set formulation and slightly improve some of the known results. In § 2 we discuss the "nonempty interior" difficulty and give an equivalent characterization. Section 3 is devoted to deducing some important properties of the (signed) distance to the fronts. In § 4 we study the nonempty interior difficulty. We give some general sufficient conditions and present some counterexamples. Section 5 provides some uniqueness properties for the distance function, which will be used in § 10. In § 6 we discuss the asymptotic limits of reaction-diffusion equations and the phase field theory. Section 7 is devoted to a formal derivation of the results. In § 8 we briefly review the theory of traveling waves of reaction-diffusion equations and we formulate our main assumptions. The main results about the phase field theory are stated in § 9; their proofs are given in § 10. Finally, in § 11 we present some possible applications and state a few open problems.

1. Geometrical evolution of level sets and degenerate parabolic pde's. In this section we recall and slightly generalize the level set formulation presented in Chen, Giga, and Goto [CGG] (see also Evans and Spruck [ESp1] for motion by mean curvature and Giga et al. [GGIS]). As mentioned in the Introduction, the underlying idea is to think of the front as the zero-level set of the solution of a pde. This type of formulation first appeared in a theoretical work of Barles [Ba1] on fronts moving with constant normal velocity. Barles [Ba1] was motivated by the computational work of Sethian [Se1] for a simple model in flame propagation. Later, Osher and Sethian [OS] extensively used this type of idea to perform numerical computations for different types of motions and, in particular, motion by mean curvature. Evans and Spruck [ESp1] provided the mathematical foundation of the level set approach for motion by mean curvature and Chen, Giga, and Goto [CGG] independently studied motions in the generality described below.

To better explain the ideas involved, we first present a formal derivation: Let Γ_t be a smooth front at time $t > 0$ and assume that $\Gamma_t = \partial D_t$, where $D_t \subset \mathbb{R}^N$ is open. The outward normal velocity V of Γ_t at $x(\in \Gamma_t)$ is given by

$$(1.1) \quad V = v(x, t, n, Dn),$$

where v is a continuous function of its arguments, n is the exterior unit normal vector

to Γ_t , and Dn is its gradient. Furthermore, we assume that there exists a smooth function $u: \mathbb{R}^N \times [0, \infty) \mapsto \mathbb{R}$ such that

$$D_t = \{x \in \mathbb{R}^N: u(x, t) > 0\}, \quad \Gamma_t = \{x \in \mathbb{R}^N: u(\cdot, t) = 0\} \quad \text{and} \quad Du \neq 0 \text{ on } \Gamma_t.$$

A classical calculation yields

$$V = \frac{u_t}{|Du|}, \quad n = -\frac{Du}{|Du|} \quad \text{and} \quad Dn = -\frac{1}{|Du|} \left(I - \frac{Du \otimes Du}{|Du|^2} \right) D^2 u.$$

Inserting the above formulae in (1.1) we obtain

$$u_t + F(x, t, Du, D^2 u) = 0,$$

where F is related to v by

$$(1.2) \quad F(x, t, p, X) = -|p|v\left(x, t, -\frac{p}{|p|}, -\frac{1}{|p|} \left(I - \frac{p \otimes p}{|p|^2} \right) X\right)$$

for $p \in \mathbb{R}^N$ and $X \in S^N$, the space of $N \times N$ matrices. An immediate consequence of (1.2) is that, for all $(x, t) \in \mathbb{R}^N \times (0, \infty)$, $p \in \mathbb{R}^N$, and $X \in S^N$, F satisfies

$$(1.3) \quad F(x, t, \lambda p, \lambda X + \mu(p \otimes p)) = \lambda F(x, t, p, X) \quad (\lambda > 0, \mu \in \mathbb{R}).$$

Any F that satisfies (1.3) will be called *geometric*.

For (1.1) to be well-posed, it is also necessary to assume that it is parabolic, i.e., that v is nonincreasing in the Dn argument. This translates in terms of (1.2) to F being (*degenerate*) *elliptic*, i.e.,

$$(1.4) \quad F(x, t, p, X) \leq F(x, t, p, Y) \quad \text{if } X \geq Y,$$

for all $(x, t) \in \mathbb{R}^N \times (0, \infty)$, $p \in \mathbb{R}^N$, and $X, Y \in S^N$. The fact that F is degenerate (in fact at least in the $p \otimes p$ direction) follows from (1.3). Finally, we point out that F is as smooth as v with a possible discontinuity at $p = 0$.

The level set approach to front propagations can be described as follows. Given a closed set Γ_0 in \mathbb{R}^N (front at time $t = 0$), choose $u_0: \mathbb{R}^N \rightarrow \mathbb{R}$ such that

$$\Gamma_0 = \{x \in \mathbb{R}^N: u_0(x) = 0\},$$

solve (in the appropriate way) the pde

$$(1.5) \quad \begin{aligned} u_t + F(x, t, Du, D^2 u) &= 0 \quad \text{in } \mathbb{R}^N \times (0, \infty), \\ u(x, 0) &= u_0(x) \quad \text{on } \mathbb{R}^N, \end{aligned}$$

and, finally, define Γ_t (the front at time t) by

$$(1.6) \quad \Gamma_t = \{x \in \mathbb{R}^N: u(x, t) = 0\}.$$

The main issues associated with such a program are (i) whether (1.5) does have a global solution allowing to define Γ_t and (ii) whether Γ_t depends only on Γ_0 and not the form of u_0 outside Γ_0 .

The first issue is settled ([ESp1], [CGG]) by considering *viscosity solutions*. Viscosity solutions, which turn out to be the correct class of generalized solutions for first- and second-order fully nonlinear pde's, were introduced by Crandall and Lions [CL] (see also [CEL] and Lions [Li] for first- and second-order equations, respectively). For the precise definition and some of the most recent developments, as well as references, we refer to the "user's guide" by Crandall, Ishii, and Lions [CIL]. In what follows (unless otherwise stated) by solution we will always mean viscosity solution. To avoid

some technicalities we will denote by (F) a set of some general assumptions needed for the statement of the next theorem. We will state and discuss these assumptions at the end of this section. Finally, we will denote by $UC(\mathcal{O})$ the set of real-valued uniformly continuous functions defined on \mathcal{O} .

THEOREM 1.1. *Assume (F), (1.3), and (1.4). Then, for any $u_0 \in UC(\mathbb{R}^N)$, there exists a unique solution $u \in UC(\mathbb{R}^N \times [0, +\infty))$ of (1.5). Moreover, if u and v are, respectively, sub- and supersolutions of (1.5) in $UC(\mathbb{R}^N \times [0, +\infty))$, then*

$$(1.7) \quad u(\cdot, 0) \leq v(\cdot, 0) \text{ in } \mathbb{R}^N \Rightarrow u \leq v \text{ in } \mathbb{R}^N \times [0, +\infty).$$

Next we discuss the issue of whether Γ_t depends only on Γ_0 . This follows from (1.3), which yields that (1.5) is invariant by nondecreasing changes $u \mapsto \psi(u)$. (See [ESp1], [CGG]).

THEOREM 1.2. *Assume the hypotheses of Theorem 1.1 hold and let $u, v \in UC(\mathbb{R}^N \times [0, +\infty))$ be solutions of (1.5) such that*

$$\begin{aligned} \{x: u(x, 0) > 0\} &= \{x: v(x, 0) > 0\}, & \{x: u(x, 0) < 0\} &= \{x: v(x, 0) < 0\}, \\ \{x: u(x, 0) = 0\} &= \{x: v(x, 0) = 0\}, \end{aligned}$$

and

$$(1.8) \quad \lim_{|x| \rightarrow +\infty} |u(x, 0)|, \lim_{|x| \rightarrow +\infty} |v(x, 0)| > 0.$$

Then, for all $t > 0$,

$$\{x: u(x, t) > 0\} = \{x: v(x, t) > 0\}, \quad \{x: u(x, t) < 0\} = \{x: v(x, t) < 0\}$$

and

$$\{x: u(x, t) = 0\} = \{x: v(x, t) = 0\}$$

This result justifies the term *equation of geometric type* for (1.5), since it yields that the evolution of the level set $\Gamma_0 \rightarrow \Gamma_t$ depends only on F and on the “signs” of the initial datum in the different regions (which in turn give a sense to the expressions “inside Γ_0 ” and “outside Γ_0 ”) and not really on the choice of the initial datum. Such a result was first obtained by Evans and Souganidis [ES1] in the case where F is independent of D^2u using representation formulae from the theory of deterministic differential games. In the generality stated above the result was obtained in [CGG]. Next we present a slightly simplified proof.

Proof. Consider the functions ϕ and ψ given by

$$\phi(t) = \inf \{v(y, 0) | u(y, 0) \geq t\} \quad \text{and} \quad \psi(t) = \sup \{v(y, 0) | u(y, 0) \leq t\}.$$

It is immediate that ϕ and ψ are nondecreasing, lower- and upper-semicontinuous (lsc and usc), respectively, and

$$(1.9) \quad \phi(u(\cdot, 0)) \leq v(\cdot, 0) \leq \psi(u(\cdot, 0)) \text{ on } \mathbb{R}^N.$$

Moreover, the assumptions on $u(\cdot, 0)$ and $v(\cdot, 0)$ yield that ϕ and ψ are actually continuous at 0 with $\phi(0) = \psi(0) = 0$. Finally, standard regularization procedures imply the existence of two sequences of nondecreasing and nonincreasing, respectively, smooth functions $(\phi_n)_n$ and $(\psi_n)_n$ such that

$$(1.10) \quad \phi = \sup_n \phi_n \quad \text{and} \quad \psi = \inf_n \psi_n.$$

Since F is geometric, $\phi_n(u)$ and $\psi_n(u)$ are solutions of (1.5). Moreover, (1.9), (1.10), and Theorem 1.1 yield

$$\phi_n(u) \leq v \leq \psi_n(u) \quad \text{in } \mathbb{R}^N \times [0, +\infty).$$

Letting $n \rightarrow \infty$ we conclude easily, since, in view of the assumptions on $u(\cdot, 0)$ and $v(\cdot, 0)$ and the definition of ϕ and ψ , $\phi(t) > 0$ if $t > 0$ and $\psi(t) < 0$ if $t < 0$. \square

We continue by discussing some examples of motions and their related “geometrical” equations.

In the first example, the hypersurface is assumed to propagate in the normal direction with velocity $v(x, t, n)$. The geometric equation in this case is

$$(1.11) \quad u_t - v \left(x, t, \frac{Du}{|Du|} \right) |Du| = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

with $F(x, t, p, M) = -v(x, t, p/|p|)|p|$ satisfying (1.3). This type of propagation, when $v \equiv c$ constant, was introduced by Landau as a flame front propagation model and was studied both analytically and numerically by Sethian [Se1] using (1.11). Then Barles [Ba1] showed the connections between (1.11) and (1.3).

Another very interesting example, both theoretically and from the applications point of view, is the motion of a hypersurface with normal velocity equal to its mean curvature. Here (1.5) takes the form

$$(1.12) \quad u_t - \Delta u + \frac{(D^2 u Du | Du)}{|Du|^2} = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

where $(\cdot | \cdot)$ denotes the usual inner product in \mathbb{R}^N . In this case (1.3) holds for every $\lambda \in \mathbb{R}$ (not only $\lambda > 0$). This yields that the equation is invariant by any change! Equation (1.12) was studied first numerically by Osher and Sethian [OS] and then analytically by Evans and Spruck [ESp1]–[ESp4] (see also Chen, Giga, and Goto [CGG], Soner [So], etc.).

Another example of propagations that arise very naturally in the theory of phase transitions is the case of anisotropic motion where (1.5) is of the form

$$(1.13) \quad u_t - |Du| \operatorname{div} \left(H \left(\frac{Du}{|Du|} \right) \right) + |Du| \beta \left(\frac{Du}{|Du|} \right) = 0$$

for some smooth functions H and β , with H convex. Equation (1.13) is studied in [So] and [CGG]. There are some very interesting models of phase transitions that yield (1.13) but with H not convex. Following a *relaxation* process, these problems give rise to (1.3) but with F discontinuous (in addition to $p = 0$) at certain directions in the gradient space. This is the subject of Gurtin, Soner, and Souganidis [GSS].

We conclude this rather long overview of the level set approach by stating and discussing assumption (F), which was necessary for the comparison result of Theorem 1.1. Assumption (F) consists of several parts, namely,

$$(x, t, p, X) \mapsto F(x, t, p, X) \text{ is bounded for bounded } (p, X)$$

$$(F_1) \quad \text{and continuous for } x \in \mathbb{R}^N, \quad t \in [0, R], \quad p \in B(0, R) \setminus \{0\}$$

$$\text{and } \|X\| \leq R, \text{ for all } R > 0.$$

$$(F_2) \quad F_*(x, t, \alpha(x-y), X) - F^*(y, t, \alpha(x-y), Y) \geq -\omega(|x-y|(1+\alpha|x-y|)),$$

where $\omega(0^+) = 0$ and for all $x, y \in \mathbb{R}^N$, $t \in (0, +\infty)$, $\alpha \geq 0$ and matrices $X, Y \in S^N$ such that $\begin{pmatrix} X & 0 \\ 0 & -Y \end{pmatrix} \leq K\alpha \begin{pmatrix} I & \\ & -I \end{pmatrix}$ for some constant $K > 0$. Finally,

$$(F_3) \quad F_*(x, t, 0, 0) = F^*(x, t, 0, 0);$$

we recall that F^* and F_* denote the upper- and lower-semicontinuous envelopes of F , respectively.

The proof of Theorem 1.1 can be found in [CGG]. The arguments of [CGG] can be, however, slightly simplified by remarking that, since (1.5) is invariant under nondecreasing changes, it is enough to have a comparison result in $BUC(\mathbb{R}^N \times [0, \infty))$, the space of bounded, uniformly continuous functions. This leads to an easier treatment of the unboundedness of the domain. In fact, with these assumptions, Theorem 1.1 extends easily to the case where either the sub- or the supersolution to be compared is discontinuous. Since we will use this remark throughout the paper, we state it as a separate theorem. (For the definition of discontinuous sub- and supersolutions we refer to [Is].)

THEOREM 1.3. *Assume (F), (1.3), and (1.4). If $u \in UC(\mathbb{R}^N \times [0, \infty))$ is a subsolution of (1.5) and $v: \mathbb{R}^N \times [0, \infty)$ is a discontinuous supersolution, then $u(\cdot, 0) \leq v(\cdot, 0)$ on \mathbb{R}^N yields $u(\cdot, t) \leq v(\cdot, t)$ on \mathbb{R}^N for all $t > 0$. A similar result holds if u is a discontinuous subsolution and $v \in UC(\mathbb{R}^N \times [0, \infty))$ is a supersolution.*

The final remark of this section is that assumption (1.8) in Theorem 1.2 can be relaxed to handle the case of unbounded fronts: we only need to assume that for each $\alpha > 0$ there exists $\varepsilon > 0$ such that

$$|u(x, 0)|, |v(x, 0)| \geq \varepsilon > 0 \quad \text{if } d(x, \Gamma_0) \geq \alpha > 0.$$

2. The nonempty interior difficulty. The level set approach seems to avoid all the geometrical difficulties related to the onset of singularities, etc. The evolution $\Gamma_0 \rightarrow \Gamma_t$ is well defined and unique. Given this fact, the next natural questions are related to the regularity of Γ_t . When $N = 2$ this issue was completely resolved by Angenent [A1], [A2] (see also the references therein). For $N \geq 3$ the issue is more complicated. In addition to a local existence result by Hamilton [H] and Evans and Spruck [ESp2] for motion by mean curvature, there are only partial regularity results (only for motion by mean curvature) due to Evans and Spruck [ESp3], [ESp4] and Ilmanen [II1], [II2].

A more basic question is whether Γ_t has an empty interior for $t > 0$. In principle, we expect Γ_t to be a hypersurface in \mathbb{R}^N ; in view of this Γ_t having interior seems rather unreasonable. This is related to the nonuniqueness features for the motion of front described by the distance function, as we will explain in the next section. Before we continue discussing this difficulty, we give a more precise definition.

DEFINITION 2.1. Let Γ_t be the evolution of Γ_0 by the level set approach. We say that $\{\Gamma_t\}_{t \geq 0}$ is regular if

$$\begin{aligned} \text{cl} \{(x, t) : u(x, t) > 0\} &= \{(x, t) : u(x, t) \geq 0\} \quad \text{and} \\ \text{int} \{(x, t) : u(x, t) \geq 0\} &= \{(x, t) : u(x, t) > 0\}. \end{aligned}$$

Clearly if $\{\Gamma_t\}_{t \geq 0}$ is regular then $\bigcup_{t \geq 0} (\Gamma_t \times \{t\})$ has an empty interior in $\mathbb{R}^N \times [0, \infty)$. Moreover in most examples the later is equivalent to Γ_t having an empty interior for all $t \geq 0$. Indeed for motion with constant normal velocity, this follows from the finite speed of propagation. For motion by mean curvature, it can be shown using explicit solutions of the form $\psi(|x|^2 + (N-1)t)$ as barriers.

We continue with a new formulation of the no empty interior question in terms of whether (1.5) has unique discontinuous solutions, with initial datum $\mathbb{1}_{\Omega_0} - \mathbb{1}_{\Omega_0^c}$, where $\mathbb{1}_A$ denotes the characteristic function of the set A , and Ω_0 and Ω_0^c are the “inside of Γ_0 ” (i.e., the set where u_0 is negative) and “outside of Γ_0 ” (i.e., the set where u_0 is positive), respectively. (See the discussion after the statement of Theorem 1.2).

THEOREM 2.1. $\{\Gamma_t\}_{t \geq 0}$ is regular if and only if there exists a unique solution of (1.5) with initial datum $\mathbb{1}_{\Omega_0} - \mathbb{1}_{\Omega_0^c}$.

In the above statement by a uniqueness we mean that if v, w are two solutions of (1.5) with the same initial data, then $(v)^* = (w)^*$ and $(v)_* = (w)_*$.

Proof of Theorem 2.1. Let $u \in UC(\mathbb{R}^N \times [0, \infty))$ be the solution of (1.5) with initial datum $d(x, \Gamma_0)$, the signed distance to Γ_0 , which is normalized to be positive inside Γ_0 and negative outside. Recall that by Theorem 1.2, it suffices to use $d(x, \Gamma_0)$ as an initial datum to obtain Γ_t . For $\varepsilon > 0$, and a scalar α set

$$u^\varepsilon(x, t) = \tanh((u(x, t) + \alpha)/\varepsilon),$$

where $\tanh(\cdot)$ is the hyperbolic tangent function. u^ε is also a solution of (1.5) (by (1.3)). The stability results for discontinuous viscosity solutions (cf. Crandall, Ishii, and Lions [CIL]) yield that the limit $u_\infty^\alpha = \lim_{\varepsilon \rightarrow 0} u^\varepsilon$ is a viscosity solution of (1.5). Moreover, the properties of \tanh yield

$$u_\infty^\alpha(x, t) = \begin{cases} 1 & \text{if } u(x, t) > \alpha, \\ -1 & \text{if } u(x, t) < \alpha, \\ 0 & \text{if } (x, t) \in \text{Int}\{u = \alpha\}. \end{cases}$$

For the rest of the points, the value of $u_\infty(x, t)$ depends on the lsc or usc envelope we consider in the definition of the discontinuous viscosity solution. Now set

$$\bar{u}_\infty = \lim_{\alpha \uparrow 0} u_\infty^\alpha \quad \text{and} \quad \underline{u}_\infty = \lim_{\alpha \downarrow 0} u_\infty^\alpha.$$

The above limits are taken in the viscosity sense (cf. [CIL]). The functions \bar{u}_∞ and \underline{u}_∞ are again solutions of (1.5). Moreover,

$$\bar{u}_\infty(x, t) = \begin{cases} 1 & \text{if } u(x, t) \geq 0 \\ -1 & \text{if } u(x, t) < 0 \end{cases} \quad \text{and} \quad \underline{u}_\infty(x, t) = \begin{cases} 1 & \text{if } u(x, t) > 0 \\ -1 & \text{if } u(x, t) \leq 0. \end{cases}$$

If $\{\Gamma_t\}_{t \geq 0}$ is not regular, \bar{u}_∞ and \underline{u}_∞ are two different discontinuous solutions of (1.5) with initial datum $\mathbb{1}_{\Omega_0} - \mathbb{1}_{\Omega_0^c}$.

Conversely, if $\{\Gamma_t\}_{t \geq 0}$ is regular, let w be a solution of (1.5) with $w(\cdot, 0) = \mathbb{1}_{\Omega_0} - \mathbb{1}_{\Omega_0^c}$ and choose a sequence $(\phi_n)_n$ of smooth functions such that $\phi_n \equiv 1$ on $[0, +\infty)$, $\phi'_n \geq 0$ in \mathbb{R} , $\phi_n(\mathbb{R}) \subset [-1, 1]$ and $\inf_n \phi_n = -1$ on $(-\infty, 0)$. Since $w^*(x, 0) \leq \phi_n(d(x, \Gamma_0))$ in \mathbb{R}^N , (1.3) and Theorem 1.3 yield $w^* \leq \phi_n(u)$ in $\mathbb{R}^N \times (0, +\infty)$ and

$$w^*(x, t) \leq -1 = \inf_n \phi_n(u(x, t)) \text{ on } \{u < 0\}.$$

On the other hand, (F_3) gives

$$F^*(x, t, 0, 0) = F_*(x, t, 0, 0) = 0;$$

hence, $+1$ and -1 are, respectively, sub- and supersolutions of (1.5). Therefore,

$$-1 \leq w_* \leq w^* \leq 1$$

and, finally, $w^* = -1$ on $\{u < 0\}$. The same method shows that $w_* = 1$ on $\{u > 0\}$, which, in view of the assumption that $\{u = 0\}$ is regular, identifies w uniquely. \square

By examining the solutions \bar{u}_∞ and \underline{u}_∞ , both equal to u_∞ in the “empty interior” case, we see that we switched from the pde formulation of the motion to a “quasi-geometric” formulation, since the notions of sub- and supersolution are only relevant on the sets $\bar{\Gamma}_t = \partial\{\bar{u}_\infty(\cdot, t) = 1\}$ and $\underline{\Gamma}_t = \partial\{\underline{u}_\infty(\cdot, t) = 1\}$. This is related to the distance function formulation for the motion, which we explain in the next section.

3. The properties of the distance function to the moving front. In this section we study the properties of the (signed) distance $d(x, \Gamma_t)$ to a front Γ_t , whose evolution has been defined by the level set approach described in § 1. The results we present here extend the work of Soner [So], who actually used the properties of the distance function to define the evolution of fronts in the case where the velocity of the front is independent of the position. Although we could do the same here, we chose not to do so, since, once the correct definition is given, all the arguments will follow exactly as in [So]. Another motivation to study the properties of the distance function, in addition to the fact that this quantity intrinsically defines the front, is that the distance function plays a central role in studying the fronts generated by reaction-diffusion equations ("phase field theory"), as we will explain in §§ 6–10.

As usual we begin with a closed set Γ_0 in \mathbb{R}^N and assign to it a notion of inside and outside in terms of the sign of its distance function. Let $\Gamma_0 \rightarrow \Gamma_t$ be the evolution of Γ_0 defined by the level set formulation. To state the main result we define the extinction time $t^* \in (0, +\infty]$ for Γ_t by

$$t^* = \sup \{t > 0 \text{ such that } \Gamma_t \neq \emptyset\}.$$

Finally, we denote by d the signed distance function to the front Γ_t .

THEOREM 3.1. *Assume that $\{\Gamma_t\}_{t \geq 0}$ is regular. Then $\underline{d} = d \wedge 0$ and $\bar{d} = d \vee 0$ satisfy, respectively,*

$$(3.1) \quad \underline{d}_t + F(x - \underline{d} D \underline{d}, t, D \underline{d}, D^2 \underline{d}) \leq 0 \quad \text{in } \mathbb{R}^N \times (0, t^*)$$

and

$$(3.2) \quad \bar{d}_t + F(x - \bar{d} D \bar{d}, t, D \bar{d}, D^2 \bar{d}) \geq 0 \quad \text{in } \mathbb{R}^N \times (0, t^*).$$

Moreover,

$$(3.3) \quad -(D^2 \underline{d} D \underline{d} | D \underline{d}) \leq 0 \quad \text{in } \{\underline{d} < 0\}$$

and

$$(3.4) \quad -(D^2 \bar{d} D \bar{d} | D \bar{d}) \geq 0 \quad \text{in } \{\bar{d} > 0\}.$$

Remark 3.2. The assumption that Γ_t has empty interior was made only to simplify the presentation. In fact, when Γ_t is not regular we can show that (3.1)–(3.4) still hold when d is replaced with appropriate functions. Indeed let $\bar{\Gamma}_t = \partial\{x: \bar{u}_\infty(x, t) = 1\}$ and $\underline{\Gamma}_t = \partial\{x: u_\infty(x, t) = 1\}$, where u_∞ and \bar{u}_∞ are defined as in the proof of Theorem 2.1. Then (3.1), (3.3) and (3.2), (3.4) hold true for $d(x, \bar{\Gamma}_t)$ and $d(x, \underline{\Gamma}_t)$. This again is related to the connections between the nonempty difficulty and the nonuniqueness in the weak geometric and distance function formulations of motions. For a detailed discussion of these connections we refer to [So].

Remark 3.3. We can read the speed of the moving front from (3.1) and (3.2). Indeed, if we know a priori that the front moves along its normal direction and if d is assumed to be smooth, then

$$d_t + F(x, t, Dd, D^2 d) = 0 \quad \text{if } d = 0,$$

which, in view of (1.1), yields $V = v(x, t, n, Dn) = -F(x, t, n, Dn)$.

Remark 3.4. We cannot expect that d will solve a pde like (1.5) as it can be observed by a direct calculation if everything is smooth. The term $x - dDd$ in (3.1) and (3.2) has a geometric meaning. Indeed, if $x \notin \Gamma_t$, then $x - dDd \in \Gamma_t$.

Proof of Theorem 3.1. We only prove (3.1) and (3.3); (3.2) and (3.4) can be obtained by similar arguments. To this end, observe that for each $k > 0$ the functions

$$w_k(x, t) = \begin{cases} 0 & \text{if } u_\infty(x, t) = 1, \\ -k & \text{if } u_\infty(x, t) = -1, \end{cases}$$

are solutions of (1.5), where $u_\infty = u_\infty^0$ is defined in the proof of Theorem 2.1. We next introduce the function

$$\bar{w}_k(x, t) = \sup_{y \in \mathbb{R}^N} \{w_k(y, t) - |x - y|\}.$$

An easy calculation yields

$$\bar{w}_k(x, t) = \max(d(x, t), -k).$$

On the other hand, standard arguments from the theory of viscosity solutions (cf. Lasry and Lions [LL], Jensen, Lions, and Souganidis [JLS]) yield that \bar{w}_k is a subsolution of (1.5). The inequalities (3.1) and (3.3) then follow easily when $d \neq 0$. If $d = 0$, we must observe that $\bar{w}_k \geq w_k$ in $\mathbb{R}^N \times (0, \infty)$ and if $\bar{w}_k(x, t) = w_k(x, t)$ at some point (x, t) , then $D^{2,+} \bar{w}_k(x, t) \subset D^{2,+} w_k(x, t)$; the last inclusion being exactly what is needed at $d = 0$. Letting $k \rightarrow \infty$ completes the proof. \square

4. When is the empty interior condition fulfilled? We hope that it has become clear by now that settling the empty interior condition is of great importance, since it may lead to some rather unintuitive situations. Unfortunately, if no conditions are imposed on Γ_0 , interior may be created for $t > 0$. See, for example, Evans and Spruck [ESp1], Soner [So], and Ilmanen [Il1] for some simple examples in this direction for motion by mean curvature. However, it can be argued that the interior in the examples of [ESp1] and [So] is due mainly to the fact that the initial data are not smooth, which, in turn, yields that the normal direction is somehow not well defined. This, of course, raises the question of finding some necessary and sufficient conditions of Γ_0 so that no interior is created. We will address this question below for the case of first-order and second-order motions whose geometric pde's are of the form

$$(4.1) \quad u_t + \alpha(x, t)|Du| = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

and

$$(4.2) \quad u_t + F(Du, D^2u) = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

with initial datum

$$(4.3) \quad u(x, 0) = d(x, \Gamma_0) \quad \text{in } \mathbb{R}^N.$$

Throughout this section we will assume that

$$(4.4) \quad \Gamma_0 = \partial\{x \in \mathbb{R}^N : d(x, \Gamma_0) < 0\} = \partial\{x \in \mathbb{R}^N : d(x, \Gamma_0) > 0\},$$

which, in particular, implies that Γ_0 has no interior.

THEOREM 4.1. *Assume (4.3), (4.4), $\alpha \in W^{1,\infty}(\mathbb{R}^N \times (0, T))$ ($\forall T > 0$), and that either (i) α does not change sign in $\mathbb{R}^N \times (0, +\infty)$ or (ii) α is independent of t . Then $\Gamma_t = \{x : u(x, t) = 0\}$ is regular, where $u \in UC(\mathbb{R}^N \times (0, \infty))$ is the solution of (4.1), (4.3). In particular Γ_t has empty interior.*

Theorem 4.1 is almost sharp. Indeed at the end of this section we will give an example of $\alpha(x, t)$ which changes sign and Γ_t develops interior. We do not, however, know whether interior is created if $\alpha \equiv \alpha(x, p/|p|)$ changes sign. (The case where $\alpha(x, p/|p|) > 0$ was treated in [Sor].)

Proof of Theorem 4.1. We present here the proof only in the case of (ii) since (i) is obtained by similar and even simpler arguments. In view of Theorem 2.1 and the discussion after Definition 2.1, it suffices to prove the uniqueness of discontinuous solutions of (4.1) with the initial datum

$$u(\cdot, 0) = \mathbb{1}_{\Omega_0} - \mathbb{1}_{\Omega_0^c} \quad \text{in } \mathbb{R}^N,$$

where $\Omega_0(x: d(x, \Gamma_0) > 0)$. To this end, we first claim that we can examine the situation separately in the sets

$$O_1 = \{x \in \mathbb{R}^N \mid \alpha(x) > 0\} \quad \text{and} \quad O_2 = \{x \in \mathbb{R}^N \mid \alpha(x) < 0\}.$$

A formal argument to understand why this claim is true consists in looking at the optimal control interpretation of (4.1) and in remarking that the paths of the dynamics starting from a point in O_1 (or O_2) can never reach the boundary of O_1 (or O_2). To justify this argument completely, we adapt some arguments introduced by Barron and Jensen [BJ1] (See also Barles [Ba2]).

Let u be a solution of (4.1) and consider the function $u^\varepsilon: O_1 \times [0, +\infty) \rightarrow \mathbb{R}$ given by

$$u^\varepsilon(x, t) = \inf_{y \in O_1} \left\{ u(y, t) + e^{-\gamma t} \frac{|x - y|^2}{\varepsilon \alpha(y)} \right\}.$$

Combining classical (in the context of viscosity solutions) computations with the arguments of [BJ1], we easily show that u^ε is an approximate subsolution of (4.1) in $O_1 \times (0, +\infty)$ for $\gamma > 0$ large enough. Moreover, u^ε is continuous and satisfies

$$u^\varepsilon(\cdot, 0) \leq \mathbb{1}_{\Omega_0} - \mathbb{1}_{\Omega_0^c} \quad \text{on } O_1.$$

If v is another solution of (4.1) and (4.3), we claim that, as $\varepsilon \rightarrow 0$,

$$u^\varepsilon \leq v_* + o(1) \quad \text{in } O_1 \times [0, +\infty).$$

Indeed, we perform the usual uniqueness arguments for viscosity solutions with a test function $\psi: (O_1 \times (0, +\infty)) \times (O_1 \times (0, +\infty)) \rightarrow \mathbb{R}$ given by

$$\psi(x, t, y, s) = u^\varepsilon(x, t) - v_*(y, s) - \frac{|x - y|^2}{\beta} - \theta \left(|x|^2 + |y|^2 + \frac{1}{\alpha(x)} + \frac{1}{\alpha(y)} \right),$$

where β and θ are small parameters. The only slight new point comes from the term

$$\left(\frac{1}{\alpha(x)} + \frac{1}{\alpha(y)} \right),$$

which takes care of the lack of boundary condition on $\partial O_1 \times (0, +\infty)$. We leave the rest of the routine but tedious details to the reader. \square

Remark 4.2. An alternative way to understand the comparison result in the proof of Theorem 4.1 is to say that (4.1) holds up to the boundary of $O_1 \times (0, +\infty)$. Indeed, let u be a usc subsolution of (4.1) and assume that $(x, t) \in \partial O_1 \times (0, +\infty)$ is a strict local maximum of $u - \phi$ for some smooth ϕ . The function

$$(y, s) \mapsto u(y, s) - \phi(y, s) - \frac{\theta}{\alpha(y)}$$

attains a maximum at $(y_\theta, s_\theta) \rightarrow (x, t)$ as $\theta \rightarrow 0$. Evaluating (4.1) at (y_θ, s_θ) and letting $\theta \rightarrow 0$ yields the result.

We next turn our attention to the case of the motion governed by (4.2); the typical example here being motion by mean curvature. We will be making the following additional assumption on F :

$$(4.5) \quad F(\mu Q'p, \mu^2 Q'XQ) = \mu^2 F(p, X)$$

for all $\mu > 0$, $p \in \mathbb{R}^n$, $X \in S^N$ and $Q \in \mathcal{O}(N)$, where Q' is the adjoint of Q and $\mathcal{O}(N)$ is the group of $N \times N$ orthogonal matrices ($Q' = Q^{-1}$).

THEOREM 4.3. *Assume that (1.3), (1.4), and (4.5) hold and that Γ_0 is of class C^2 . In addition, assume that there exist nonnegative constants c_i ($i = 1, 2, 3$), a skewsymmetric matrix H , and $x_0 \in \mathbb{R}^N$ such that*

$$(4.6) \quad c_1(x - x_0) \cdot Dd(x) + c_2 H(x - x_0) \cdot Dd(x) - c_3 F(Dd(x), D^2 s(x)) \neq 0 \text{ on } \Gamma_0,$$

where d is the signed distance to Γ_0 . Then the set $\bigcup_{t>0} (\Gamma_t \times \{t\})$ has empty interior in $\mathbb{R}^N \times (0, +\infty)$.

The left-hand side of (4.6) is the generator of rotation, dilations, and translations in (x, t) evaluated at $t = 0$ on Γ_0 . Condition (4.6) includes as special cases results of Ilmanen [Il1] and Soner [So] for motion by mean curvature. On the other hand, (4.6) is not necessary. Indeed, recent work of Soner and Souganidis [SS] (see also Altschuler, Angenent, and Giga [AAG]) for bodies of rotation moving by mean curvature shows that there exist smooth Γ_0 's which do not satisfy (4.6), but their evolution never develops interior. It follows, however, that (4.6) holds near the singularities of Γ_t [SS]. This is related to a conjecture of DeGiorgi [D]. A related observation is that if (4.6) holds at a later time, this again yields no interior. For the case of mean curvature, Evans and Spruck [ESp4] also showed that under some assumptions on Γ_0 , almost every level set of the solution of (1.12) does not develop interior. Finally, at the end of this section we give an example where interior is created if the velocity depends on t .

Proof of Theorem 4.3. Let $u \in UC(\mathbb{R}^N \times (0, \infty))$ be the unique solution of (4.2) and (4.3) and, for $h > 0$, define the function

$$u_h(x, t) = \Phi(u((1 + c_1 h) e^{c_2 h H}(x - x_0) + x_0, (1 + c_1 h)t + c_3 h)),$$

where Φ is some increasing smooth function with $\Phi(0) = 0$ to be chosen later. In view of (1.3) and (4.5), u_h is also a solution of (4.2), since H being skewsymmetric yields $Q = e^{c_2 h H} \in \mathcal{O}(N)$. Moreover, if h is small enough, there exists some $\eta > 0$ such that

$$(4.7) \quad |u(\cdot, 0) - u_h(\cdot, 0)| \geq \eta h \quad \text{on } \mathbb{R}^N.$$

Assuming for the moment (4.7), we observe that Theorem 1.1 yields either $u_h \leq u - \eta h$ or $u_h \leq u + \eta h$ in $\mathbb{R}^N \times (0, \infty)$. If $\bigcup_{t>0} (\Gamma_t \times \{t\})$ has interior, either of the above inequalities, however, yields a contradiction, for if $u = 0$ in some neighborhood of a point (x_0, t_0) , then so does u_h for h sufficiently small.

We return now to the proof of (4.7). We first observe that we may choose Φ so that we only need to check (4.7) in a small neighborhood of Γ_0 . But for a suitable choice of such a neighborhood u is smooth. We can therefore perform the expansion

$$\begin{aligned} u((1 + c_1 h) e^{c_2 h H}(x - x_0) + x_0, c_3 h) &= u(x, 0) + h(c_1(x - x_0) \cdot Du(x, 0) \\ &\quad + c_2 H(x - x_0) \cdot Du(x, 0) + c_3 u_t(x, 0)) + o(h). \end{aligned}$$

Using (4.6), that $u(x, 0) = d(x)$, and the fact that the equation holds for small $t > 0$ (since Γ_0 is smooth) we conclude the proof. \square

In fact, with a modification of the above proof, we can prove that Γ_t is regular. We leave this modification to the reader.

We continue with an example of interior for a motion governed by (4.1).

PROPOSITION 4.4. *Consider (4.1) in $\mathbb{R} \times (0, \infty)$ with $\alpha(x, t) = x - t$. There exists an interval $I = (\beta, \gamma)$ such that the evolution $\Gamma_0 \rightarrow \Gamma_t$ has nonempty interior at some $t_0 > 0$, where $\Gamma_0 = \partial I$.*

Proof. In view of Theorem 2.1, it suffices to show that there exists I such that the equation

$$(4.8) \quad \begin{cases} u_t + (x - t)|u_x| = 0 & \text{in } \mathbb{R} \times (0, \infty), \\ u(x, 0) = (\mathbb{1}_I - \mathbb{1}_{I^c})(x) & \text{on } \mathbb{R}, \end{cases}$$

has more than one solution. To this end, choose $x_0 > 0$, solve the forward and backward ordinary differential equations (ode's)

$$\dot{X}_{\pm}(t) = \pm \alpha(X_{\pm}(t), t) \quad \text{with } X_{\pm}(x_0) = x_0,$$

and set $\beta = X_+(0)$, $\eta = X_-(0)$, and $I = (\beta, \eta)$. We will compute the minimal and maximal solution of (4.8), using the control interpretation of this equation. Indeed, consider the dynamics given by

$$\dot{y}_x(s) = \alpha(y_x(s), s)v(s), \quad y_x(t) = x,$$

where $v(\cdot) \in L^\infty((0, +\infty), [-1, 1])$ is the control process. Following Barles and Perthame [BaP] or Barron and Jensen [BJ2], we can prove easily that the minimal and maximal solution of $u_t + (x - t)|u_x| = 0$ in $\Omega_1 = \{x > t\}$ are, respectively,

$$u_*(x, t) = \inf_{v(\cdot)} u_*(y_x(0), 0) \quad \text{and} \quad u^*(x, t) = \inf_{v(\cdot)} u^*(y_x(0), 0),$$

where $u(x, 0) = (\mathbb{1}_I - \mathbb{1}_{I^c})(x)$. It is easy to see from the above formulae that $u_* \equiv -1$ on $\{(x, t): x = t\}$, $u^* = -1$ on $\{(x, t): x = t\} \setminus \{(x_0, x_0)\}$ and $u^*(x_0, x_0) = 1$. We now turn our attention to $\Omega_2 = \{(x, t): x < t\}$. Here the maximal and minimal solutions are, given by, respectively,

$$\bar{u}(x, t) = \sup_{v(\cdot)} \{-\mathbb{1}_{\{\tau=t\}} + u^*(y_x(\tau), \tau)\mathbb{1}_{\{\tau>t\}}\}$$

and

$$\underline{u}(x, t) = \sup_{v(\cdot)} \{-\mathbb{1}_{\{\tau=t\}} + u_*(y_x(\tau), \tau)\mathbb{1}_{\{\tau>t\}}\},$$

where, for each $v(\cdot)$, τ is the exit time from Ω_2 . It follows that, while $\underline{u} \equiv -1$ in Ω_2 , \bar{u} equals 1 at each point $(x, t) \in \Omega$ for which the trajectory y_x may reach the point (x_0, x_0) . It is easy to check that the set of these points is exactly the region $\{(x, t) \in \Omega_2: X_+(t) \leq x \leq X_-(t)\}$ which has a nonempty interior.

Since (4.8) has a nonuniqueness feature, we conclude by Theorem 2.1. \square

The next example of nonuniqueness corresponds to volume preserving mean curvature flow. The derivation of this motion and its significance for applications is discussed in § 11.

Let Γ_0 be the union of three disjoint circles in \mathbb{R}^2 , i.e., $\Gamma_0 = \partial B(x_1, R_0) \cup \partial B(x_2, R_0) \cup \partial B(x_3, r_0)$, with $x_i \in \mathbb{R}^2 (i = 1, 2, 3)$ to be chosen later and $0 < r_0 < R_0$. We consider the motion of Γ_0 with normal velocity

$$V = -\operatorname{div}(Dn) + \alpha(t) \quad (t > 0),$$

where $\alpha(t) = 2\pi N(t)L^{-1}(t)$, $N(t)$ and $L(t)$ being the number of disjoint parts of Γ_t and its length, respectively. In view of this explicit formula, at least for small time,

$$\Gamma_t = \partial B(x_1, R_t) \cup \partial B(x_2, R_t) \cup \partial B(x_3, r_t),$$

where R_t, r_t satisfy the ode's

$$\dot{R}_t = -R_t^{-1} + \alpha(t) \quad \text{and} \quad \dot{r}_t = -r_t^{-1} + \alpha(t) \quad \text{with} \quad \alpha(t) = 3(2R_t + r_t)^{-1}.$$

Let $t_1 = \sup \{t > 0 \text{ such that } r_t > 0\}$. The form of Γ_t above is valid for all $t \in (0, t_1)$. Since t_1 is independent of the choice of the x_i 's we can choose x_1 and x_2 so that $|x_1 - x_2| = 2R_{t_1}$. In view of this choice,

$$\Gamma_{t_1} = \partial B(x_1, R_{t_1}) \cup \partial B(x_2, R_{t_1}),$$

with the two circles touching at a point. There are two possible evolutions for $t \geq t_1$ depending on whether we think of Γ_t as one set or two separate ones. In the first case Γ_t moves with $\alpha(t) = 2\pi (\text{length}(\Gamma_t))^{-1}$ and actually converges to $\partial B((x_1 + x_2)/2, R_\infty)$, as $t \rightarrow \infty$, where $R_\infty = (2R_0^2 + r_0^2)^{1/2}$. In the second case, Γ_t remains stationary (i.e., $\alpha(t) \equiv R_{t_1}^{-1}$ for $t > t_1$).

We conclude the discussion about the “nonempty interior” difficulty with a general comment for the t -dependent velocities. It appears that we cannot hope to have a general theorem guaranteeing no interior without making very severe restrictions on the t -dependence of the normal velocity. The reason for this claim is the following. In principle, all motions have some “pathological” situations, where interior develops. We can take any such motion, perturb its velocity by a time dependent forcing term so that to drive the front to the pathological situation, and then simply turn off the time.

5. Uniqueness results for the distance function formulation. As mentioned in § 3, we can have a weak formulation of the propagation of a front in terms of whether the signed distance to the front satisfies the inequalities (3.1) and (3.2). A natural question to ask is whether (3.1) and (3.2) are enough to identify the distance function uniquely, i.e., if z satisfies (3.1) and (3.2) and $z(x, 0) = d(x, \Gamma_0)$, is it true that $z \equiv d$? In addition to being a natural mathematical question to ask, having such information simplifies a lot some of the analysis of the “phase field” theory.

In the following, and only to considerably simplify the presentation, we will only consider the equation

$$(5.1) \quad u_t - \theta \left(\Delta u - \frac{(D^2 u Du | Du)}{|Du|^2} \right) + \alpha(x, t)|Du| = 0 \text{ in } \mathbb{R}^N \times (0, \infty)$$

with the initial datum

$$(5.2) \quad u(x, 0) = d(x, \Gamma_0) \quad \text{in } \mathbb{R}^N,$$

with $\theta \geq 0$ and $\alpha \in W^{1,\infty}(\mathbb{R}^N \times (0, \infty))$. (Some of the arguments and the conclusions below hold if $\theta = \theta(x, t)$ (under some assumptions) as well as for anisotropic motions. We will discuss these situations elsewhere.)

As before, we denote by $\Gamma_t = \{x: u(x, t) = 0\}$. Theorem 3.1 and the discussion following it say that the functions $d_1 = d(x, \bar{\Gamma}_t)$ and $d_2 = d(x, \underline{\Gamma}_t)$ (where $\bar{\Gamma}_t = \partial\{x: u(x, t) > 0\}$ and $\underline{\Gamma}_t = \partial\{x: u(x, t) \geq 0\}$) satisfy the inequalities

$$(5.3) \quad z_t - \theta \Delta z + \alpha(x - zDz, t) \leq 0, \quad 1 - |Dz| = 0 \text{ in } \{z < 0\}$$

and

$$(5.4) \quad z_t - \theta \Delta z + \alpha(x - zDz, t) \geq 0, \quad |Dz| - 1 = 0 \text{ in } \{z > 0\}.$$

Of course, if the no-interior condition holds for every $t > 0$, (5.3) and (5.4) are satisfied by $d = d(x, \Gamma_t)$. The inequalities in (5.3) and (5.4) are a combination of (3.1) and (3.3) and (3.2) and (3.4), respectively, as they apply to (5.1). On the other hand, the equalities in (5.3) and (5.4) follow from the differentiability properties of the distance function and the definition of viscosity solutions.

Next we look into the converse of Theorem 3.1, i.e., we are interested in whether (5.3) and (5.4) identify z as the distance function.

THEOREM 5.1. *If the usc (respectively lsc) function z satisfies (5.2) and (5.3) (respectively (5.2) and (5.4)), then*

$$(5.5) \quad z \leq d_2 \quad \text{in } \{z < 0\} \supset \{d_2 < 0\},$$

$$(5.6) \quad z \geq d_1 \quad \text{in } \{z > 0\} \supset \{d_1 > 0\},$$

respectively. If z satisfies (5.2)–(5.4) and $\{\Gamma_t\}_{t \geq 0}$ is regular, then

$$(5.7) \quad z(x, t) = d(x, \Gamma_t) \quad \text{in } \mathbb{R}^N \times [0, \infty).$$

Proof. The proof is based on the following two lemmas.

LEMMA 5.2. *If z is usc (respectively lsc) and satisfies (5.3) (respectively, (5.4)), then z is a subsolution (respectively supersolution) of*

$$(5.8) \quad z_t - \theta \left(\Delta z - \frac{(D^2 z D z | D z)}{|D z|^2} \right) + \alpha(x - z D z, t) |D z| = 0$$

in $\{z < 0\}$ (respectively, $\{z > 0\}$).

LEMMA 5.3. *If a usc (respectively lsc) function z satisfies (5.3) (respectively (5.4)), then for C large enough, $\underline{z} = e^{Ct}(z \wedge 0)$ (respectively $\bar{z} = e^{Ct}(z \vee 0)$) is a subsolution (respectively supersolution) of (5.1).*

We first conclude the proof of the theorem and then prove the lemmas. We proceed by proving (5.5), since (5.6) follows in a similar way. To this end, observe that, since \underline{z} (defined in Lemma 5.3) is a subsolution of (5.1), Theorem 1.2 yields $\underline{z} \leq u \wedge 0$ in $\mathbb{R}^N \times (0, \infty)$; recall that $u \wedge 0$ is still a solution of (5.1), since $\Phi(u) = u \wedge 0$ is an increasing change of u . So, if $u < 0$ (or, equivalently, if $d_2 < 0$), $z < 0$ and the proof of (5.5) is complete.

Finally, if $\{\Gamma_t\}_{t \geq 0}$ is regular, then $d_1 = d_2 = d$ and (5.5) and (5.6) yield

$$\{z < 0\} = \{d < 0\}, \quad \{z > 0\} = \{d > 0\} \quad \text{and} \quad \{z = 0\} = \{d = 0\};$$

therefore, $z = d$ by the uniqueness results for the equations $|Dz| - 1 = 0$ and $1 - |Dz| = 0$, respectively, in $\{z > 0\} = \{d > 0\}$ and $\{z < 0\} = \{d < 0\}$. \square

We now return to the proofs of the lemmas.

Proof of Lemma 5.2. We only treat the case of a usc z that satisfies (5.6); the other case is proved similarly. Since z is usc, the set $\Omega = \{z < 0\}$ is open. Moreover, z being a solution of $1 - |Dz| = 0$ in $\Omega_t = \{x: z(x, t) < 0\}$ for all $t > 0$, yields

$$z(x, t) = \sup \{z^*(y, t) - |x - y|: y \in \Omega_t\}, \quad \forall x \in \Omega_t,$$

where $z^*(y, t) = \limsup_{\Omega_t \ni y' \rightarrow y} z(y', t)$. This formula implies that z is locally semiconvex with respect to x , i.e. $\partial^2 z / \partial \chi^2 \geq -C$ in Ω , for all unit vectors $\chi \in \mathbb{R}^N$. Next we define the ε -supconvolution z^ε of z in Ω with respect to t by

$$z^\varepsilon(x, t) = \sup_{(x, s) \in \Omega} \left\{ z(x, s) - \frac{(t - s)^2}{\varepsilon} \right\}.$$

It follows easily that, for (x, t) belonging to compact subset V of Ω and $\varepsilon > 0$ small enough, the supremum is actually achieved in Ω (and not on $\partial\Omega$) and that z_ε satisfies

$$(5.9) \quad 1 - |Dz^\varepsilon| = 0 \quad \text{and} \quad z_t^\varepsilon - \theta \Delta z^\varepsilon + \alpha(x - z^\varepsilon Dz^\varepsilon, t) \leq C\varepsilon \quad \text{in } V,$$

where C depends only on the Lipschitz bound of α . Let $(x_0, t_0) \in \Omega$ be a strict local maximum of $z - \phi$ in Ω for smooth ϕ and take $V \Subset \Omega$ in (5.9) to be a neighborhood of (x_0, t_0) . Since $z^\varepsilon \rightarrow z$, there exists $(x_\varepsilon, t_\varepsilon) \in V$ maximum points of $z^\varepsilon - \phi$, such that $(x_\varepsilon, t_\varepsilon) \rightarrow (x_0, t_0)$ as $\varepsilon \rightarrow 0$. Now we use Alexandrov's Maximum Principle-type arguments, brought in the theory of viscosity solutions by Jensen [J]. More precisely, Lemma A.3 of [CIL] implies the existence of $X_\varepsilon \in S^N$ such that

$$(5.10) \quad (\phi_t(x_\varepsilon, t_\varepsilon), D\phi(x_\varepsilon, t_\varepsilon), X_\varepsilon) \in J^{2,+} z^\varepsilon(x_\varepsilon, t_\varepsilon) \quad \text{and} \quad -K \leq X_\varepsilon \leq D_{xx}^2 \phi(x_\varepsilon, t_\varepsilon),$$

for some constant K , which is related to semiconvexity constant of z and, therefore, of z^ε in V ; the upper bound on X_ε comes from the Maximum Principle. (We refer to [CIL] for the definition of $J^{2,+}$.) Also we claim that $X_\varepsilon D\phi(x_\varepsilon, t_\varepsilon) = 0$. Indeed, since $|Dz^\varepsilon| = 1$ almost everywhere, $D_{xx}^2 z^\varepsilon Dz^\varepsilon = 0$ at any point where z^ε is twice differentiable. On the other hand (cf. [CIL, Lemma A.3]), $X_\varepsilon D\phi(x_\varepsilon, t_\varepsilon)$ is obtained as a limit of $D_{xx}^2 z^\varepsilon Dz^\varepsilon$ evaluated at nearby points. Finally, recall that $D\phi(x_\varepsilon, t_\varepsilon) = Dz^\varepsilon(x_\varepsilon, t_\varepsilon)$, since z^ε is differentiable at maximum points of $z^\varepsilon - \phi$ (again due to the semiconvexity).

Inserting all the information in (5.9) we obtain

$$\begin{aligned} & \phi_t - \theta \left(\Delta \phi - \frac{(D^2 \phi D\phi | D\phi)}{|D\phi|^2} \right) + \alpha(x_\varepsilon - z^\varepsilon D\phi, t_\varepsilon) \\ & \leq \phi_t - \theta \left(\text{Tr}(X_\varepsilon) - \frac{(X_\varepsilon D\phi | D\phi)}{|D\phi|^2} \right) + \alpha(x_\varepsilon - z^\varepsilon D\phi, t_\varepsilon) \leq C\varepsilon, \end{aligned}$$

where in the two inequalities above, z^ε and ϕ and its derivatives are evaluated at $(x_\varepsilon, t_\varepsilon)$. Letting $\varepsilon \rightarrow 0$ we conclude, the proof. \square

Proof of Lemma 5.3. We again only present the proof in the case that z is a usc subsolution.

If c is larger than the Lipschitz constant of α , Lemma 5.2 implies that $e^{ct}z$ is a subsolution of (5.1), since $|Dz| = 1$ yields

$$\alpha(x - zDz, t) \geq \alpha(x, t) - cz = \alpha(x, t)|Dz| - cz.$$

To conclude let $(\psi_n)_n$ be a sequence of smooth functions such that $\psi_n(t) = 0$ if $t \geq -1/n$, $\psi_n' \geq 0$ and $\psi_n' \rightarrow 1$ uniformly on compact subsets of $(-\infty, 0]$. Using the preceding lemma, it is easy to check that $\psi_n(e^{ct}z)$ is a subsolution of (5.1). Letting $n \rightarrow \infty$ we conclude, since $\psi_n(e^{ct}z) \rightarrow e^{ct}(z \wedge 0)$. \square

6. Asymptotic limits of Reaction-Diffusion equations-Phase field theory. Reaction-diffusion equations of the form

$$(6.1) \quad \phi_t - \Delta \phi + f(x, t, \phi) = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

arise naturally in many areas of applications, such as phase transitions, flame propagations, pattern formations, chemical kinetics, etc. In most of these applications, fronts develop for large times as the boundaries of the regions where the solution ϕ of (6.1) converges to the different equilibria of the vector field f (cf. Fife [Fi]). For a discussion of some cases where the solutions of (6.1) converge to the different equilibria of f we refer to Aronson, and Weinberger [ArW], Fife and McCleod [FiM], etc. The main issue is to identify the rate at which ϕ converges to the different equilibria. For this,

we must have a better understanding of the fronts and, in particular, the way they propagate. In the case $f(x, t, \phi) = f(\phi)$, formal results of Fife [Fi] and Caginalp [Ca1]–[Ca3] imply that the fronts propagate with normal velocity

$$(6.2) \quad V = \alpha + \frac{1}{t}\kappa + O\left(\frac{1}{t^2}\right) \quad (t \gg 1),$$

when κ denotes the curvature.

Our goal here is to justify (6.2) rigorously in the generality of (6.1). One way to do this is to scale ϕ so as to capture the different terms in the asymptotic expansion (6.2). To obtain the first term, the appropriate scaling is $(x/\varepsilon, t/\varepsilon)$. If $\alpha = 0$, we then go to the next scaling $(x/\varepsilon, t/\varepsilon^2)$. These considerations give rise to singular perturbation problems of the form

$$(6.3) \quad \phi_t^\varepsilon - \varepsilon \Delta \phi^\varepsilon + \frac{1}{\varepsilon} f^\varepsilon(x, t, \phi^\varepsilon) = 0 \quad \text{in } \mathbb{R}^N = (0, +\infty)$$

and

$$(6.4) \quad \phi_t^\varepsilon - \Delta \phi^\varepsilon + \frac{1}{\varepsilon^2} f^\varepsilon(x, t, \phi^\varepsilon) \times 0 \quad \text{in } \mathbb{R}^N \times (0, +\infty),$$

with initial data

$$(6.5) \quad \phi^\varepsilon(\cdot, 0) = \phi_0^\varepsilon(\cdot) \quad \text{on } \mathbb{R}^N.$$

Here ϕ_0^ε is a given function that initializes the front and f^ε is some approximation of f . Singular perturbation problems of the form (6.3) and (6.4) are of independent interest for they also arise in models with slow diffusion and fast reaction, in phase transitions, etc.

In the following we study the behavior, as $\varepsilon \rightarrow 0$, of (6.3) and (6.4) under the assumption that $\phi \mapsto f^\varepsilon(x, t, \phi)$ is a “cubic-type” nonlinearity, i.e., it has two stable and one unstable equilibria. Typical examples of f^ε are

$$(6.6) \quad f^\varepsilon(x, t, q) = 2(q - \varepsilon\mu(x, t))(q^2 - 1),$$

$$(6.7) \quad f^\varepsilon(x, t, q) = 2(q - \mu(x, t))(q^2 - 1),$$

and

$$(6.8) \quad f^\varepsilon(x, t, q) = 2(q - \mu)(q^2 - 1) + \varepsilon\theta^\varepsilon(x, t),$$

where $\theta^\varepsilon, \mu \in W^{1,\infty}(\mathbb{R}^n \times [0, +\infty))$ are given and μ takes values in $(-1, 1)$.

To simplify the presentation, we restrict ourselves to problems where the second-order operator is the Laplacian, although all the arguments can be modified to apply to more general elliptic operators (under, of course, suitable hypotheses). This will be addressed in the future. Finally, we remark that the case where f^ε is of “quadratic” type (i.e., f^ε has one stable and one unstable equilibria) has been studied by probabilistic methods by Freidlin [Fr] and, in greater generality, by pde-type techniques by Evans and Souganidis [ES2], [ES3] and Barles, Evans, and Souganidis [BaES]. The latter work actually studies a general system of reaction-diffusion equations.

We conclude this section with a brief discussion of the “phase field” approach to study propagating fronts. This consists of first studying the behavior of ϕ^ε as $\varepsilon \rightarrow 0$ in (6.3) and (6.4) and then defining the propagating front as the boundary of the regions where the ϕ^ε ’s converge to the different equilibria of the vector field. The advantage of this approach, which is rather indirect, is that it avoids any discussion

of the empty interior and the nonuniqueness difficulties at least at first glance provided of course that such a convergence can be proved. However, it will become apparent below that the convergence is closely related to the interior issue. Perhaps another advantage of the phase field approach is that it allows other numerical methods. This way to study motion by mean curvature was proposed by Bronsard and Kohn [BrK] and DeGiorgi [D]. A byproduct of our analysis in the following sections is that the phase field formulation is equivalent to the level set and distance function ones, taking into account the nonempty interior difficulty.

7. Formal discussion. In this section we discuss, in a formal way, the essential mathematical difficulties involved in the study of (6.3) and (6.4). To simplify the arguments, we consider the special case

$$(7.1) \quad f^\varepsilon(x, t, q) = f_0(q) - \varepsilon\theta = 2(q - \mu)(q^2 - 1) - \varepsilon\theta \quad (\theta \in \mathbb{R}).$$

We begin observing that, for sufficiently small $\varepsilon > 0$, there exists $h_-^\varepsilon(\theta) < h_0^\varepsilon(\theta) < h_+^\varepsilon(\theta)$ such that

$$f^\varepsilon(x, t, h_-^\varepsilon(\theta)) = f^\varepsilon(x, t, h_0^\varepsilon(\theta)) = f^\varepsilon(x, t, h_+^\varepsilon(\theta)) = 0.$$

Set

$$(7.2) \quad \begin{aligned} m^\varepsilon(\theta) &= h_+^\varepsilon(\theta) - h_-^\varepsilon(\theta), \\ q^\varepsilon(r, \theta) &= h_-^\varepsilon(\theta) + m^\varepsilon(\theta)(1 + \exp(-m^\varepsilon(\theta)[r + r^\varepsilon(\theta)]))^{-1} (r \in \mathbb{R}), \\ c^\varepsilon(\theta) &= 2h_0^\varepsilon(\theta) - h_+^\varepsilon(\theta) - h_-^\varepsilon(\theta), \end{aligned}$$

where $r^\varepsilon(\theta)$ is chosen so that $q^\varepsilon(0, \theta) = h_0^\varepsilon(\theta)$. A straightforward calculation yields

$$(7.3) \quad q_{rr}^\varepsilon + c^\varepsilon(\theta)q_r^\varepsilon = f_0(q^\varepsilon) - \varepsilon\theta$$

with

$$(7.4) \quad \lim_{r \rightarrow \pm\infty} q^\varepsilon(r, \theta) = h_\pm^\varepsilon(\theta);$$

in other words, q^ε is the *traveling wave* corresponding to the nonlinearity $f_0 - \varepsilon\theta$, which travels with speed $c^\varepsilon(\theta)$. Indeed, if we set

$$\Phi^\varepsilon(\xi, t) = q^\varepsilon(\xi - c^\varepsilon(\theta)t) \quad \text{in } \mathbb{R} \times (0, \infty),$$

then

$$\Phi_t - \Phi_{\xi\xi} = f_0(\Phi) - \varepsilon\theta \quad \text{in } \mathbb{R} \times (0, \infty).$$

In fact, for any “cubic-type” nonlinearity, there exists a unique pair of traveling wave and speed satisfying (7.3) and (7.4). A detailed discussion of this fact as well as references will be given in the next section.

We now return to (6.3) and write the solution ϕ^ε as

$$\phi^\varepsilon = q^\varepsilon\left(\frac{z^\varepsilon}{\varepsilon}, \theta\right) \quad \text{in } \mathbb{R}^N \times (0, \infty).$$

A simple calculation yields

$$\frac{1}{\varepsilon}q_r^\varepsilon[z_t^\varepsilon - \varepsilon\Delta z^\varepsilon + c^\varepsilon(\theta)] - \frac{1}{\varepsilon}q_{rr}^\varepsilon(|Dz^\varepsilon|^2 - 1) = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

where q_r^ε and q_{rr}^ε are evaluated at $(z^\varepsilon/\varepsilon, \theta)$.

Analyzing the two terms in the above equation separately, as $\varepsilon \rightarrow 0$, we formally conclude that $|Dz^\varepsilon| \cong 1$ and, therefore,

$$z^\varepsilon(x, t) \cong \text{signed distance function of } x \text{ to } \Gamma_t,$$

where Γ_t is the interface, and

$$z_t^\varepsilon - \varepsilon \Delta z^\varepsilon + c^\varepsilon(\theta) \cong 0 \text{ on } \Gamma_t.$$

Since $h_0^\varepsilon(\theta) \cong \mu + \varepsilon \theta (f_0'(\mu))^{-1}$ and $h_\pm^\varepsilon(\theta) \cong \pm 1 + \varepsilon \theta (f_0'(\pm 1))^{-1}$, (7.2) yields

$$\lim_{\varepsilon \rightarrow 0} c^\varepsilon(\theta) = 2\mu.$$

Therefore, always formally, Γ_t moves with normal velocity $V = -2\mu$. The geometric pde that gives Γ_t as the zero level set of its solutions is

$$u_t + 2\mu|Du| = 0 \text{ in } \mathbb{R}^N \times (0, \infty).$$

In view of the discussion in § 6, to consider (6.4) with the vector field f^ε given by (7.1), we must assume $\mu = 0$, i.e.,

$$f^\varepsilon(x, t, q) = 2q(q^2 - 1) - \varepsilon\theta.$$

Proceeding as for (6.3) above, we write

$$\phi^\varepsilon = q\left(\frac{z^\varepsilon}{\varepsilon}, \theta\right) \text{ in } \mathbb{R}^N \times (0, \infty)$$

and find

$$\frac{1}{\varepsilon} q_r^\varepsilon [z_t^\varepsilon - \Delta z^\varepsilon + \varepsilon^{-1} c^\varepsilon(\theta)] - \frac{1}{\varepsilon^2} q_{rr}^\varepsilon (|Dz^\varepsilon|^2 - 1) = 0 \text{ in } \mathbb{R}^N \times (0, \infty),$$

where q_r^ε and q_{rr}^ε are evaluated at $(z^\varepsilon/\varepsilon, \theta)$. Arguing as before, we find (formally) that $z^\varepsilon(x, t) \cong \text{signed distance function from } x \text{ to } \Gamma_t$, where Γ_t is the interface, and

$$z_t^\varepsilon - \Delta z^\varepsilon + \varepsilon^{-1} c^\varepsilon(\theta) \cong 0 \text{ on } \Gamma_t.$$

Using the expressions for $h_0^\varepsilon(\theta)$, $h_\pm^\varepsilon(\theta)$ and (7.2) we find

$$\lim_{\varepsilon \rightarrow 0} \varepsilon^{-1} c^\varepsilon(\theta) = -\frac{3}{2}\theta.$$

Therefore, always formally, Γ_t moves with normal velocity

$$V = \text{mean curvature} + \frac{3}{2}\theta.$$

The corresponding geometric pde is

$$u_t - \left(\Delta u - \frac{(D^2 u Du | Du)}{|Du|^2} \right) - \frac{3}{2}\theta|Du| = 0 \text{ in } \mathbb{R}^N \times (0, \infty).$$

8. Traveling waves. Here we discuss the existence and the general properties of traveling waves for functions $u \mapsto f^\varepsilon(x, t, u)$, which have the property that, for a and ε small, the function $u \mapsto f^\varepsilon(x, t, u) - \varepsilon a$ behaves like a “cubic function” of u . More precisely, we assume that, for a and ε sufficiently small, the equation $f^\varepsilon(x, t, u) - \varepsilon a = 0$ has exactly three zeros: $h_-^\varepsilon(x, t, a) < h_0^\varepsilon(x, t, a) < h_+^\varepsilon(x, t, a)$. Moreover, we assume that

$$(8.1) \quad \begin{aligned} f^\varepsilon(x, t, \cdot) - \varepsilon a &> 0 && \text{in } (h_-^\varepsilon, h_0^\varepsilon) \cup (h_+^\varepsilon, +\infty), \\ f^\varepsilon(x, t, \cdot) - \varepsilon a &< 0 && \text{in } (-\infty, h_-^\varepsilon) \cup (h_0^\varepsilon, h_+^\varepsilon), \\ f_u^\varepsilon(x, t, h_\pm^\varepsilon) &\cong \gamma > 0, \end{aligned}$$

with γ independent of (x, t, a, ε) .

Since, for fixed (x, t, a, ε) , the function $u \mapsto f^\varepsilon(x, t, u) - \varepsilon a$ satisfies the hypotheses of Aronson and Weinberger [ArW] and Fife and McLeod [FiM], there exists a unique pair $(q^\varepsilon(r, x, t, a), c^\varepsilon(x, t, a))$ such that

$$(8.2) \quad q_{rr}^\varepsilon(r, x, t, a) + c^\varepsilon(x, t, a)q_r^\varepsilon(r, x, t, a) = f^\varepsilon(x, t, q^\varepsilon(r, x, t, a)) - \varepsilon a$$

and

$$(8.3) \quad \lim_{r \rightarrow \pm\infty} q^\varepsilon(r, x, t, a) = h_\pm^\varepsilon(x, t, a) \quad \text{and} \quad q^\varepsilon(0, x, t, a) = h_0^\varepsilon(x, t, a);$$

the second part of (8.3) is necessary to fix q^ε since (8.2) is invariant under translation in r .

We continue listing a set of technical assumptions that we will be making on $(q^\varepsilon, c^\varepsilon)$. We then verify these assumptions for a particular class of f^ε 's, which arise naturally in applications. To this end, we assume that, as $\varepsilon \rightarrow 0$,

$$(8.4) \quad q^\varepsilon \text{ and } c^\varepsilon \text{ depend smoothly on } (x, t, a),$$

$$(8.5) \quad h_\pm^\varepsilon(x, t, a) \rightarrow h_\pm(x, t, a), \quad h_0^\varepsilon(x, t, a) \rightarrow h_0(x, t, a),$$

and either

$$(8.6) \quad c^\varepsilon(x, t, a) \rightarrow \alpha(x, t, a)$$

or

$$(8.7) \quad -\varepsilon^{-1}c^\varepsilon(x, t, a) \rightarrow \alpha(x, t, a), \quad \text{if } c^\varepsilon(x, t, a) \rightarrow 0,$$

with all the limits local uniform in (x, t, a) . Moreover, if

$$\alpha(x, t) = \alpha(x, t, 0), \quad h_\pm(x, t) = h_\pm(x, t, 0), \quad \text{and} \quad h_0(x, t) = h_0(x, t, 0),$$

we assume that there exists $K > 0$, independent of (x, t) , such that, for ε and a small enough and all (x, t) ,

$$(8.8) \quad |\alpha(x, t) - \alpha(y, t)| \leq K|x - y|.$$

If (8.7) holds, we also assume

$$(8.9) \quad \begin{aligned} & \text{(i)} \quad |h_{\pm t} - \Delta h_\pm| \leq K \\ & \text{(ii)} \quad \lim_{\varepsilon \rightarrow 0} \sup_{(x, t, r, a)} [\varepsilon |q_t^\varepsilon| + \varepsilon |\Delta q^\varepsilon| + |Dq_r^\varepsilon|] = 0 \end{aligned}$$

$$\text{(iii)} \quad \frac{1}{\varepsilon} |q_{rr}^\varepsilon(r, x, t, a)| + \frac{1}{\varepsilon^2} |q_r^\varepsilon(r, x, t, a)| \leq K e^{-K\delta/\varepsilon} \quad \text{for all } |r| \geq \delta.$$

Finally, for all (x, t) and ε, a sufficiently small, we assume

$$(8.10) \quad q_r^\varepsilon \geq 0 \quad \text{and} \quad q_a^\varepsilon \geq 0.$$

Next we present an example where the above hypotheses hold true. Indeed, consider

$$(8.11) \quad f^\varepsilon(x, t, q) = 2(q - \mu^\varepsilon(x, t))(q^2 - 1) - \varepsilon \theta^\varepsilon(x, t),$$

where $\theta^\varepsilon: \mathbb{R}^N \times (0, \infty) \rightarrow \mathbb{R}$ is a given function. Let $h_0^\varepsilon, h_\pm^\varepsilon, q^\varepsilon$ and c^ε be as in § 7 for each (x, t) and define

$$\begin{aligned} h_0^\varepsilon(x, t, a) &= h_0^\varepsilon(\theta^\varepsilon(x, t) + a), & h_\pm^\varepsilon(x, t, a) &= h_\pm^\varepsilon(\theta^\varepsilon(x, t) + a), \\ q^\varepsilon(r, x, t, a) &= q^\varepsilon(r, \theta^\varepsilon(x, t) + a), \end{aligned}$$

and

$$c^\varepsilon(x, t, a) = c^\varepsilon(\theta^\varepsilon(x, t) + a).$$

It is immediate that (8.4) holds (if θ^ε is smooth) and that (8.5) holds with $h_\pm(x, t, a) = \pm 1$ and $h_0(x, t, a) = \mu$; (8.6) holds with $\alpha(x, t, a) = 2\mu(x, t)$ where $\mu = \lim_{\varepsilon \rightarrow 0} \mu^\varepsilon$. If $\mu(x, t) = 0$, then (8.7) yields $\alpha(x, t, a) = \frac{3}{2}(\theta(x, t) + a)$, provided that $\theta^\varepsilon(x, t) \rightarrow \theta(x, t)$ uniformly. In view of the above, (8.8) needs

$$(8.12) \quad |\theta(x, t) - \theta(y, t)| \leq K|x - y| \quad \text{or} \quad |\mu(x, t) - \mu(y, t)| \leq K|x - y|.$$

To conclude, using the explicit formulae in (7.2) we compute

$$D_t q^\varepsilon = q_\theta^\varepsilon \theta_t^\varepsilon, \quad D_x q^\varepsilon = q_\theta^\varepsilon D_x \theta^\varepsilon, \quad \Delta_x q^\varepsilon = q_\theta^\varepsilon \Delta_x \theta^\varepsilon + q_{\theta\theta}^\varepsilon |D_x \theta^\varepsilon|^2.$$

Since $|q_\theta| \leq \varepsilon K$ and $|q_{\theta\theta}| \leq \varepsilon^2 K$ for some $K > 0$, (8.9) (ii) holds if θ^ε is such that

$$(8.13) \quad \lim_{\varepsilon \rightarrow 0} \varepsilon [\sup_{(x,t)} (\varepsilon |\theta_t^\varepsilon| + \varepsilon |\Delta \theta^\varepsilon| + |D\theta^\varepsilon|)] = 0.$$

For (8.12) and (8.13) to hold, it suffices to assume that

$$(8.14) \quad (\theta^\varepsilon)_{\varepsilon > 0} \text{ is uniformly bounded in } C^{2,1}(\mathbb{R}^N \times [0, \infty)).$$

Finally, (8.9)(iii) and (8.11) hold provided $4\varepsilon|\theta^\varepsilon| \leq 1$, which follows from (8.14) for ε small.

We conclude this section observing that similar computations are possible for

$$(8.15) \quad f^\varepsilon(x, t, q) = 2(\theta^\varepsilon(x, t)q - \mu^\varepsilon(x, t))((\theta^\varepsilon(x, t)q)^2 - 1).$$

9. Asymptotic behavior of reaction-diffusion equations; the main results. We next state our main theorem about the behavior of the solution ϕ^ε of (6.3) and (6.4). To study (6.3) we consider f^ε 's that satisfy (8.1)–(8.6) and (8.8) and (8.10). For (6.4) we will consider f^ε 's such that (8.1)–(8.5) and (8.7)–(8.10) hold. In either case, we will denote by $(q^\varepsilon(r, x, t), c^\varepsilon(x, t))$ the pair of traveling wave and speed which corresponds to f^ε and we will assume

$$(9.1) \quad \alpha(x, t, a) \geq \alpha(x, t) \quad \text{for all } a > 0.$$

Throughout the discussion below we will be assuming that

$$(9.2) \quad \phi^\varepsilon(x, 0) = q^\varepsilon\left(\frac{d(x, \Gamma_0)}{\varepsilon}, x, 0\right) \quad \text{on } \mathbb{R}^N,$$

where Γ_0 is a closed set in \mathbb{R}^N . The last assumption can be weakened at the expense of rather lengthy arguments. We will address this issue elsewhere.

In view of the (formal) discussion in § 7 we expect that the limiting behavior of ϕ^ε will be governed by the geometric pde's

$$(9.3) \quad u_t - \alpha(x, t)|Du| = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

for (6.3) and

$$(9.4) \quad u_t - \left(\Delta u - \frac{(D^2 u Du | Du)}{|Du|^2} \right) - \alpha(x, t)|Du| = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

for (6.4), with (by (9.2)),

$$(9.5) \quad u(x, 0) = d(x, \Gamma_0) \quad \text{on } \mathbb{R}^N.$$

THEOREM 9.1. *Let ϕ^ε be the solution of (6.3), (9.2) with f^ε satisfying (8.1)–(8.6) and (8.8), (8.10), and (9.1). If u is the solution of (9.3), (9.5), then, as $\varepsilon \rightarrow 0$,*

$$(9.6) \quad \begin{aligned} \text{(i)} \quad & \phi^\varepsilon(x, t) \rightarrow h_+(x, t) \quad \text{if } u(x, t) > 0, \\ \text{(ii)} \quad & \phi^\varepsilon(x, t) \rightarrow h_-(x, t) \quad \text{if } u(x, t) < 0, \end{aligned}$$

with the limits locally uniform in $\{(x, t) : u(x, t) \neq 0\}$.

THEOREM 9.2. *Let ϕ^ε be the solution of (6.4), (9.2) with f^ε satisfying (8.1)–(8.5), (8.7)–(8.10), and (9.1). If u is the solution of (9.4), (9.5), then, locally uniformly in $\{(x, t) : u(x, t) \neq 0\}$, as $\varepsilon \rightarrow 0$*

$$(9.7) \quad \begin{aligned} & \text{(i)} \quad \phi^\varepsilon(x, t) \rightarrow h_+(x, t) \quad \text{if } u(x, t) > 0, \\ & \text{(ii)} \quad \phi^\varepsilon(x, t) \rightarrow h_-(x, t) \quad \text{if } u(x, t) < 0. \end{aligned}$$

In the special case where $f^\varepsilon(x, y, u) = 2(u - \mu)(1 - u^2)$, Barles, Bronsard, and Souganidis [BaBS] studied the limiting behavior of the solutions ϕ^ε of (6.3). Gärtner [G] also studied the same problem when $f^\varepsilon(x, t, u) = f(x, t, u)$ by a combination of probabilistic and analytic techniques. Evans, Soner, and Souganidis [ESS] studied the limiting behavior of ϕ^ε in (6.4) when $f(u) = 2u(1 - u^2)$; this problem was first studied in the context of radially symmetric functions by Bronsard and Kohn [BrK]. Finally, Chen [Ch] and DeMottoni and Shatzman [DS] obtained results similar to Theorems 9.1 and 9.2 (for special cases of f) assuming, however, that Γ_t is a smooth surface. No such assumption is made here.

We conclude this section by remarking that we can actually obtain more precise results than (9.6) and (9.7). Indeed, it is possible to obtain WKB-type expressions for ϕ^ε of the form

$$\phi^\varepsilon(x, t) = q^\varepsilon\left(\frac{d(x, \Gamma_t) + o(1)}{\varepsilon}, x, t\right).$$

This is done in § 10.1 for some simple cases. The arguments for the general case are, however, rather complicated and will be presented elsewhere.

10. Proofs. Instead of presenting a general proof for Theorems 9.1 and 9.2, we will first give some less general but more direct arguments utilizing the results of § 5. At the end we will turn to the general case. The reason for doing this is that in the less general cases it is possible to work directly at the $\varepsilon = 0$ level, as opposed to the general case where we need to build super- and subsolutions for $\varepsilon > 0$. The latter approach ties us down to cases where the maximum principle holds.

10.1 The (x, t) -independent case. Here we assume that f^ε (and therefore q^ε) is independent of (x, t) and is given by (6.6) for (6.4) and (6.7) for (6.3). In fact, the traveling wave in either case is $q(r) = \tanh(r)$ ($r \in \mathbb{R}$) and the speed $2\varepsilon\mu$ and 2μ for (6.6) and (6.7), respectively.

Following the discussion in § 7, if

$$(10.1) \quad \phi^\varepsilon = q\left(\frac{z^\varepsilon}{\varepsilon}\right) \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

then z^ε solves

$$(10.2) \quad z_t^\varepsilon - \varepsilon \Delta z^\varepsilon + 2q\left(\frac{z^\varepsilon}{\varepsilon}\right)(|Dz^\varepsilon|^2 - 1) + 2\mu = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

in the case of (6.3), and

$$(10.3) \quad z_t^\varepsilon - \Delta z^\varepsilon + \frac{2}{\varepsilon}q\left(\frac{z^\varepsilon}{\varepsilon}\right)(|Dz^\varepsilon|^2 - 1) + 2\mu = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

in the case of (6.4), with, in either case,

$$(10.4) \quad z^\varepsilon(x, 0) = d(x, \Gamma_0) \quad \text{on } \mathbb{R}^N.$$

We want to study the behavior of z^ε as $\varepsilon \rightarrow 0$. To this end, we assume for the moment that $(z^\varepsilon)_{\varepsilon > 0}$ is locally uniformly bounded in $\mathbb{R}^N \times (0, T)$ for some $T > 0$ and we proceed. Since (10.2) and (10.3) are translation invariant with respect to x , it is immediate that

$$(10.5) \quad |Dz^\varepsilon| \leq 1 \quad \text{in } \mathbb{R}^N \times (0, \infty).$$

On the other hand, the form of (10.2) and (10.3) makes any kind of estimate z_t^ε hopeless. To circumvent this difficulty we use the, by now classical, half-relaxed limit techniques described in Barles and Perthame [BaP] (see also [CIL]), i.e., we consider the functions

$$(10.6) \quad z^*(x, t) = \limsup_{\substack{\varepsilon \rightarrow 0 \\ s \rightarrow t}} z^\varepsilon(x, s) \quad \text{and} \quad z_*(x, t) = \liminf_{\substack{\varepsilon \rightarrow 0 \\ s \rightarrow t}} z^\varepsilon(x, s).$$

We begin with (10.3), which can be rewritten as

$$(10.7) \quad z_t^\varepsilon - \Delta z^\varepsilon + 2\mu = -\frac{2}{\varepsilon} q\left(\frac{z^\varepsilon}{\varepsilon}\right) (|Dz^\varepsilon|^2 - 1).$$

The form of q and (10.5) yield

$$-\frac{2}{\varepsilon} q\left(\frac{z^\varepsilon}{\varepsilon}\right) (|Dz^\varepsilon|^2 - 1) \geq 0 \quad \text{if } z^\varepsilon > 0$$

and

$$-\frac{2}{\varepsilon} q\left(\frac{z^\varepsilon}{\varepsilon}\right) (|Dz^\varepsilon|^2 - 1) \leq 0 \quad \text{if } z^\varepsilon < 0.$$

Using (10.5), (10.7), and the above inequalities we get that z^* is a usc subsolution of (5.3) with $\alpha = -2\mu$ and a solution of $1 - |Dz| = 0$ in $\{z < 0\}$ and that z_* is an lsc supersolution of (5.4) with $\alpha = -2\mu$ and a solution of $|Dz| - 1 = 0$ in $\{z > 0\}$. That z^* (respectively, z_*) is a subsolution (respectively, supersolution) of (5.3) in $\{z < 0\}$ (respectively, (5.4) in $\{z > 0\}$) follows from (10.7) and the above inequalities, that z^* (respectively, z_*) solves $1 - |Dz| = 0$ (respectively, $|Dz| - 1 = 0$) in $\{z < 0\}$ (respectively, $\{z > 0\}$) follows the passage to the limit in both (10.5) and (10.7).

Theorem 5.1 implies that $z^* \leq d_2$ in $\{z^* < 0\} \supset \{u < 0\}$ and $z_* \geq d_1$ in $\{z_* > 0\} \supset \{u > 0\}$, where u is the solution of (9.4) with $\alpha = -2\mu$. Moreover if the “empty interior” condition holds, Theorem 5.1 yields $z^*(\cdot, t) = z_*(\cdot, t) = d(\cdot, \Gamma_t)$; therefore we have the result.

In the case of (10.2), we rewrite the equation as

$$z_t^\varepsilon - \varepsilon \Delta z^\varepsilon + 2\mu |Dz^\varepsilon| = -(|Dz^\varepsilon| - 1) \left(2\mu + 2q\left(\frac{z^\varepsilon}{\varepsilon}\right) (|Dz^\varepsilon| + 1) \right)$$

and pass to the limit using sign-type arguments, similar to the first case but for the limiting equation. Indeed, since $\mu \in (-1, 1)$, we obtain

$$z_t + 2\mu |Dz| \leq 0 \quad \text{in } \{z < 0\} \quad \text{and} \quad z_t + 2\mu |Dz| \geq 0 \quad \text{in } \{z > 0\},$$

where above we have suppressed the z^* and z_* notation. The arguments of the proofs of Theorem 5.1 and Lemmas 5.2 and 5.3 yield

$$\{z^* < 0\} \supset \{u < 0\} \quad \text{and} \quad \{z_* > 0\} \supset \{u > 0\};$$

we conclude as before.

It remains to prove the uniform local bound on z^ε . Such a bound is easy for (10.2) and we leave it up to the reader; here we concentrate on (10.3). Let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ be a C^2 function such that $\psi \equiv 0$ in $[0, +\infty)$ and $\psi(-\infty) = -1$ with $\psi' > 0$ in $(-\infty, 0)$ and ψ''

bounded and consider the function $\bar{\omega}_\varepsilon$ defined by $\bar{\omega}_\varepsilon = \psi(z^\varepsilon)$. In view of the choice of ψ , it is clear that $-1 \leq \bar{\omega}_\varepsilon \leq 0$, i.e., $\bar{\omega}_\varepsilon$ is bounded. Next we define

$$\bar{w}^*(x, t) = \limsup_{\substack{\varepsilon \rightarrow 0 \\ s \rightarrow t}} \bar{\omega}_\varepsilon(x, s);$$

\bar{w}^* is well defined and $\bar{w}^* = -1$ if $z^* = -\infty$, $\bar{w}^* = \psi(z^*)$ if $z^* \in (-\infty, 0)$ and $\bar{w}^* = 0$ if $z^* \geq 0$. Combining the above with arguments from the proof of Theorem 5.1, it can be shown that \bar{w}^* is a subsolution of the two-sided variational inequality

$$\max \left(w, \min \left(w + 1, w_t - \left(\Delta w - \frac{(D^2 w D w | D w)}{|D w|^2} \right) + 2\mu |D w| \right) \right) = 0.$$

A direct modification of the usual comparison results yields

$$\bar{w}^* \leq \psi(u) \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

where u is the solution of (9.4) with $\alpha = -2\mu$. Arguing in exactly the same way with $w_\varepsilon = -\psi(-z_\varepsilon)$, we find

$$w_*(x, t) = \lim_{\substack{\varepsilon \rightarrow 0 \\ s \rightarrow t}} w_\varepsilon(x, t) \geq -\psi(-u) \quad \text{in } \mathbb{R}^N \times (0, \infty).$$

We conclude as follows: Let $t^* = \sup \{t > 0: \text{there exists } x \in \mathbb{R}^N \text{ such that } u(x, t) > 0\}$. If $t < t^*$, the sets $\{u > 0\}$ and $\{u < 0\}$ and, therefore, $\{z^* < 0\}$ and $\{z_* < 0\}$ are nonempty. Then there exist points in a bounded region of \mathbb{R}^N (depending only on u) such that $z^* < 0$ and $z_* > 0$. The local uniform bound then follows from (10.5) for all $T < t^*$. If $t > t^*$, then $z^* < 0$ and therefore $\phi_\varepsilon \rightarrow -1$ at any such point. \square

10.2. The (x, t) -dependent case. We now study (6.3) and (6.4) in the case where f^ε is given by (6.6)–(6.8). We only give the proof for (6.4) for f^ε given by (6.6); the other cases can be treated similarly. First, we recall that the traveling wave q^ε associated with (6.6) is still $q^\varepsilon = q = \tanh$. As before we perform the change $\phi^\varepsilon = q(z^\varepsilon/\varepsilon)$ and find (10.7) takes the form

$$(10.7) \quad z_t^\varepsilon - \Delta z^\varepsilon + 2\mu(x, t) + \frac{2}{\varepsilon} q\left(\frac{z^\varepsilon}{\varepsilon}\right) (|Dz^\varepsilon|^2 - 1) = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty)$$

with

$$z^\varepsilon(x, 0) = d(x, \Gamma_0) \quad \text{on } \mathbb{R}^N.$$

The main difference between this case and the (x, t) -independent one is that (10.7) is no longer translation invariant with respect to x . Instead of (10.5) here we have

$$(10.8) \quad |Dz^\varepsilon| \leq e^{Ct} \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

where C is the Lipschitz constant of 2μ with respect to x .

Next we introduce the function $\underline{z}^\varepsilon$ defined by

$$\underline{z}^\varepsilon(x, t) = \inf_{y \in \mathbb{R}^N} (\eta(z^\varepsilon(y, t)) + |x - y|),$$

where $\eta \in C^2$ is such that: $\eta(0) = 0$, $\eta' > 1$ in $(0, \infty)$, $\eta' < 1$ in $(-\infty, 0)$ and $\beta > \eta'' > 0$ and $\eta \geq -\beta^{-1}$ on \mathbb{R} for some $\beta > 0$. Since η is bounded from below, it is clear that the infimum in the definition of $\underline{z}^\varepsilon$ is achieved for some $y^\varepsilon(x)$; y^ε also depends on t

but we suppress this here. We now perform the usual arguments for this type of inf-convolution. If $y^\varepsilon(x) \neq x$, then

$$Dz^\varepsilon(y^\varepsilon(x), t) = \frac{1}{\eta'(z^\varepsilon(y^\varepsilon(x), t))} \frac{x - y^\varepsilon(x)}{|x - y^\varepsilon(x)|},$$

hence

$$(10.9) \quad q\left(\frac{z^\varepsilon}{\varepsilon}\right)(|Dz^\varepsilon|^2 - 1) = q\left(\frac{z^\varepsilon}{\varepsilon}\right)\left(\frac{1}{\eta'(z^\varepsilon)^2} - 1\right) \leq 0 \text{ at } (y^\varepsilon(x), t).$$

On the other hand, if $y^\varepsilon(x) = x$ and $\underline{z}^\varepsilon(x, t) = z^\varepsilon(x, t) > 0$, then $|Dz^\varepsilon(x, t)| \leq 1$ and

$$(10.10) \quad q\left(\frac{z^\varepsilon}{\varepsilon}\right)(|Dz^\varepsilon|^2 - 1) \leq 0.$$

Combining the last two inequalities and (10.8) we obtain

$$(10.11) \quad \underline{z}_t^\varepsilon - \Delta \underline{z}^\varepsilon + \beta e^{2Ct} + 2\eta'(z^\varepsilon(y^\varepsilon(x), t))\mu(y^\varepsilon(x), t) \geq 0 \text{ in } \{\underline{z}^\varepsilon > 0\}.$$

As in the previous section we assume that the z^ε 's (and therefore the $\underline{z}^\varepsilon$'s) are locally uniformly bounded in $\mathbb{R}^N \times (0, \infty)$ and we consider

$$z_*(x, t) = \liminf_{\substack{\varepsilon \rightarrow 0 \\ s \rightarrow t}} z^\varepsilon(x, s) \quad \text{and} \quad \underline{z}(x, t) = \liminf_{\substack{\varepsilon \rightarrow 0 \\ s \rightarrow t}} \underline{z}^\varepsilon(x, s).$$

Letting $\varepsilon \rightarrow 0$ in (10.7) we get

$$(10.12) \quad \text{sgn}(z_*)(|Dz_*| - 1) \geq 0 \text{ in } \mathbb{R}^N \times (0, \infty).$$

We also must send $\varepsilon \rightarrow 0$ in (10.11). To do so we assume that $y^\varepsilon(x) \rightarrow y(x)$ for some $y(x)$ as $\varepsilon \rightarrow 0$ (since the family $(y^\varepsilon(x))_\varepsilon$ is bounded, $y^{\varepsilon_n}(x) \rightarrow y(x)$ for some $y(x)$ at least along some subsequence); hence

$$\underline{z}_t - \Delta \underline{z} + \beta e^{2Ct} + 2\eta'(z_*(y(x), t))\mu(y(x), t) \geq 0 \text{ in } \{\underline{z} > 0\}.$$

Now we remark that

$$\underline{z}(x, t) = \eta(z_*(y(x), t)) + |x - y(x)|;$$

the definitions of z^ε and \underline{z} together with (10.9), (10.10), and (10.12) and the properties of η yield $z_*(y(x), t) = 0$ and, therefore, $\underline{z}(x, t) = d(x, \{z_* = 0\})$ and $\eta'(z_*(y(x), t)) = 1$. We conclude by combining the arguments of the previous section and the ones of the proof of Theorem 5.1 and letting $\beta \rightarrow 0$.

10.3. The general case. Unfortunately, we cannot prove Theorems 9.1 and 9.2 in the case of general f^ε by a direct passage to the limit; one of the main difficulties being the lack of an explicit formula for the traveling wave q^ε and its speed c^ε . Here we will proceed by constructing sub- and supersolutions for (6.3) and (6.4) following ideas introduced in Evans, Soner, and Souganidis [ESS]. As before, we will only present the proof of Theorem 9.2; Theorem 9.1 is proved in a similar way with some modifications noted below. We begin with some preliminary facts.

For fixed $\delta, a > 0$, let $u^{\delta, a}$ be the solution of

$$(10.13) \quad \begin{aligned} u_t^{\delta, a} + F(x, t, Du^{\delta, a}, D^2 u^{\delta, a}) &= (\alpha(x, t, a) - \alpha(x, t))|Du^{\delta, a}| \\ u^{\delta, a}(x, 0) &= d(x, \Gamma_0) + \delta \quad \text{on } \mathbb{R}^N \end{aligned}$$

where

$$F(x, t, p, X) = -\operatorname{tr}(X) + \frac{(Xp|p)}{|p|^2} - \alpha(x, t)|p|.$$

If

$$d^{\delta,a}(x, t) = d(x, \{y: u^{\delta,a}(y, t) = 0\}),$$

Theorem 3.1 yields that

$$(10.14) \quad d_t^{\delta,a} - \Delta d^{\delta,a} - \alpha(x - d^{\delta,a} Dd^{\delta,a}, t, a) \geq 0 \text{ in } \{d^{\delta,a} > 0\}.$$

Following the proof of Lemma 3.1 of [ESS], we define

$$(10.15) \quad w^{\delta,a}(x, t) = \eta_\delta(d^{\delta,a}(x, t)),$$

where, as in [ESS], $\eta_\delta: \mathbb{R} \rightarrow \mathbb{R}$ is a smooth function satisfying

$$(10.16) \quad \begin{aligned} \eta_\delta(z) &= -\delta \text{ if } z \leq -\frac{\delta}{4}, \\ \eta_\delta(z) &= z - \delta \text{ if } z \geq \frac{\delta}{2}, \\ \eta_\delta(z) &\leq -\frac{\delta}{2} \text{ if } z \leq \frac{\delta}{2}, \\ 0 &\leq \eta'_\delta \leq C \text{ and } |\eta''_\delta| \leq C\delta^{-1} \text{ on } \mathbb{R}, \end{aligned}$$

where $C > 0$ is independent of δ . A straightforward modification of Lemma 3.1 of [ESS] together with (10.15) yields the following lemma.

LEMMA 10.1. *There exists a constant C , independent of δ and a , such that*

$$(10.17i) \quad w_t^{\delta,a} - \Delta w^{\delta,a} - \alpha(x, t, a) |Dw^{\delta,a}| \geq -\frac{C}{\delta} \quad \text{in } \mathbb{R}^N \times [0, t^*),$$

$$(10.17ii) \quad w_t^{\delta,a} - \Delta w^{\delta,a} - \alpha(x - w^{\delta,a} Dw^{\delta,a}, t, a) \geq 0 \quad \text{on } \left\{d^{\delta,a} > \frac{\delta}{2}\right\},$$

and

$$(10.18) \quad |Dw^{\delta,a}| = 1 \text{ in } \left\{d^{\delta,a} > \frac{\delta}{2}\right\},$$

where t^* is the extinction time of $\{u^{\delta,a} = 0\}$.

Finally, we define

$$(10.19) \quad \Phi^\varepsilon(x, t) = q^\varepsilon\left(\frac{w^{\delta,a}(x, t)}{\varepsilon}, x, t, a\right) \quad \text{on } \mathbb{R}^N \times [0, \infty),$$

where, for notational simplicity, we do not exhibit the dependence of Φ^ε on δ and a .

PROPOSITION 10.2. *Assume that f^ε satisfies the hypotheses of Theorem 9.2. Then, for every $a > 0$, Φ^ε is a supersolution of (6.4), if $\varepsilon \leq \varepsilon_0(\delta, a)$ and $\delta \leq \delta_0(a)$.*

The proof of the proposition is similar to the proof of Theorem 3.2 of [ESS]. The form, however, of Φ^ε is different than the one used in [ESS]. As usual we will present the proof as if $w^{\delta,a}$ has actual derivatives, keeping in mind that everything actually has to be checked in the viscosity sense; we will leave it to the reader to do so.

Proof. We must show that

$$(10.20) \quad \Phi_t^\varepsilon - \Delta \Phi^\varepsilon + \frac{1}{\varepsilon^2} f^\varepsilon(x, t, \Phi^\varepsilon) \geq 0$$

for $\varepsilon \leq \varepsilon_0(\delta, a)$ and $\delta \leq \delta_0(a)$. Using the equation for $q^\varepsilon(\xi, x, t, a)$, we calculate

$$(10.21) \quad \begin{aligned} \Phi_t^\varepsilon - \Delta \Phi^\varepsilon + \frac{1}{\varepsilon^2} f^\varepsilon(x, t, \Phi^\varepsilon) &= J^\varepsilon - \frac{q_{rr}^\varepsilon}{\varepsilon^2} (|Dw^{\delta,a}|^2 - 1) \\ &\quad + \frac{1}{\varepsilon} q_r^\varepsilon \left(w_t^{\delta,a} - \Delta w^{\sigma,a} + \frac{c^\varepsilon}{\varepsilon} \right) + \frac{a}{\varepsilon}, \end{aligned}$$

where q_r^ε and q_{rr}^ε are evaluated at $(w^{\delta,a}/\varepsilon, x, t, a)$, with

$$(10.22) \quad J^\varepsilon(x, t) = \left(q_t^\varepsilon + \frac{2}{\varepsilon} Dq_r^\varepsilon Dw^{\delta,a} + \Delta q^\varepsilon \right) \left(\frac{w^{\delta,a}}{\varepsilon}, x, t, a \right).$$

In view of its definition, it is immediate that $|Dw^{\delta,a}| \leq C$ where C is as in (10.16). Therefore, by (8.9) (ii),

$$(10.23) \quad J^\varepsilon = \frac{o(1)}{\varepsilon} \text{ as } \varepsilon \rightarrow 0 \text{ uniformly in } (x, t, \delta, a).$$

We proceed by examining three cases.

Case 1. $\delta/2 < d^{\delta,a} < 2\delta$.

Using (10.18), the Lipschitz continuity of α with respect to x , the fact that $d^{\delta,a} < 2\delta$, and the form of η_δ , we get

$$w_t^{\delta,a} - \Delta w^{\delta,a} - \alpha(x, t, a) \geq -C\delta \quad \text{and} \quad |Dw^{\delta,a}| = 1.$$

Substituting in (10.22) and employing (10.23) we obtain

$$(10.24) \quad \Phi_t^\varepsilon - \Delta \Phi^\varepsilon + \frac{1}{\varepsilon^2} f^\varepsilon(x, t, \Phi^\varepsilon) \geq \frac{1}{\varepsilon} \left[q_r^\varepsilon \left(-C\delta + \frac{c^\varepsilon(x, t, a)}{\varepsilon} + \alpha(x, t, a) \right) + a + o(1) \right],$$

where again q_r^ε is evaluated at $(w^{\delta,a}/\varepsilon, x, t, a)$. Since $\varepsilon^{-1}c^\varepsilon(x, t, a) \rightarrow -\alpha(x, t, a)$ as $\varepsilon \rightarrow 0$, uniformly in (x, t, a) , we see that the right side of (10.24) is positive if ε and δ are sufficiently small.

Case 2. $d^{\delta,a} \leq \delta/2$.

In this case the choice of η_δ yields

$$w^{\delta,a} \leq -\delta/2.$$

Consequently, (8.9) (iii) yields that

$$\frac{1}{\varepsilon} \left| q_r^\varepsilon \left(\frac{w^{\delta,a}}{\varepsilon}, x, t, a \right) \right| + \frac{1}{\varepsilon^2} \left| q_{rr}^\varepsilon \left(\frac{w^{\delta,a}}{\varepsilon}, x, t, a \right) \right| \leq Ke^{-K\delta/\varepsilon}$$

for some appropriate constant K . Using that $|Dw^{\delta,a}| \leq C$ as well as (10.17) in (10.21)

we obtain

$$\Phi_t^\varepsilon - \Delta \Phi^\varepsilon + \frac{1}{\varepsilon^2} f^\varepsilon(x, t, \Phi^\varepsilon) \geq K e^{-K\delta/\varepsilon} \left[-\frac{c}{\delta} - c \right] + o(1) + \frac{a}{\varepsilon}, \quad \text{as } \varepsilon \rightarrow 0;$$

for ε small enough the right hand side of the above inequality is again positive.

Case 3. $d^{\delta,a} > 2\delta$.

In this case we have $w^{\delta,a} > \delta$. Using (10.17) and (8.9) (iii) we conclude as in the previous case. \square

We are now ready to give the proof of Theorem 9.2.

Proof of Theorem 9.2. Fix $(x_0, t_0) \in \mathbb{R}^N \times [0, t^*)$ such that $u(x_0, t_0) = -\beta < 0$. The stability of solutions of the geometric pde's yields that $u^{\delta,a} \rightarrow u$, as $\delta, a \rightarrow 0$, uniformly in (x, t) . We choose, therefore, sufficiently small a and δ so that

$$(10.25) \quad u^{\delta,a}(x_0, t_0) < -\frac{\beta}{2} < 0.$$

Let Φ^ε be given by (10.19). In addition to being a supersolution of (6.4) for sufficiently small $\varepsilon > 0$, Φ^ε satisfies

$$\Phi^\varepsilon(x, 0) \geq q^\varepsilon\left(\frac{d(x, \Gamma_0)}{\varepsilon}, x, 0\right) \quad \text{on } \mathbb{R}^N,$$

where the last inequality follows from the fact that

$$w^{\delta,a}(x, 0) = \eta_\delta(d(x, \Gamma_0) + \delta) \geq d(x, \Gamma_0).$$

It follows by the standard comparison theorem for viscosity solutions and (8.10) that

$$\Phi^\varepsilon \leq \phi^\varepsilon \quad \text{in } \mathbb{R}^N \times [0, t^*).$$

On the other hand, (10.25) yields $d^{\delta,a}(x_0, t_0) < 0$; hence

$$\limsup_{\varepsilon \rightarrow 0} \phi^\varepsilon(x_0, t_0) \leq \limsup_{\varepsilon \rightarrow 0} \Phi^\varepsilon(x_0, t_0) = h_-(x_0, t_0).$$

To prove the reverse inequality, we consider $\hat{\Phi}(x, t) = h_-(x, t) - \gamma$ for some $\gamma > 0$. Since $h_- \in C^{2,1}$,

$$\frac{\partial \hat{\Phi}}{\partial t} - \Delta \hat{\Phi} + \frac{1}{\varepsilon^2} f^\varepsilon(x, t, \hat{\Phi}) \leq K + \frac{1}{\varepsilon^2} [-\gamma f_q^\varepsilon(x, t, h_-(x, t)) + o(\gamma)].$$

By (8.1), the right-hand side is negative for small ε and γ . Hence by the maximum principle

$$\liminf_{\varepsilon \rightarrow 0} \phi^\varepsilon(x, t) \geq h_-(x, t) - \gamma \quad \text{for all } (x, t) \text{ and } \gamma > 0.$$

We conclude by letting $\gamma \rightarrow 0$. A simple modification of the above arguments yields that $\phi^\varepsilon \rightarrow h_-$ locally uniformly in $\{u < 0\}$.

The fact that $\phi^\varepsilon \rightarrow h_+$ in $\{u > 0\}$ follows in a similar way, provided we construct a subsolution of (6.4).

To prove Theorem 9.1 we must consider the traveling waves associated by $f^\varepsilon - a$ and argue about a lower bound on $-\varepsilon \Delta w^{a,\delta}$. The latter follows from the facts that $w^{a,\delta} \neq 0$ if and only if $d^{a,\delta} \geq \delta/4$ and $\Delta d^{a,\delta} \geq -C/d^{a,\delta}$ in $\{d^{a,\delta} > 0\}$. \square

11. Possible applications. In this section we briefly discuss two applications where (6.4) arises naturally, with f^ε of the form

$$(11.1) \quad f^\varepsilon(x, t, q) = 2q(q^2 - 1) - \varepsilon \theta^\varepsilon(x, t),$$

which, in view of the discussion in § 8, satisfies the desired properties, provided $(\theta^\varepsilon)_\varepsilon$ is bounded in $C^{2,1}$. On the other hand, we do not know whether $(\theta^\varepsilon)_\varepsilon$ satisfies this necessary condition.

Example 1 (volume constraint). Let Ω be a bounded domain in \mathbb{R}^N with an outward normal vector $n(x)$, $x \in \partial\Omega$ and consider the reaction-diffusion equation

$$(11.2) \quad \phi_t^\varepsilon - \Delta \phi^\varepsilon + \frac{2}{\varepsilon^2} \phi^\varepsilon ((\phi^\varepsilon)^2 - 1) = a^\varepsilon(t) \quad \text{in } \Omega,$$

$$\frac{\partial \phi^\varepsilon}{\partial n} = 0 \quad \text{on } \partial\Omega,$$

where

$$(11.3) \quad a^\varepsilon(t) = \lambda^\varepsilon(\phi^\varepsilon(\cdot, t)) = \frac{1}{\varepsilon^2} \frac{1}{\text{meas}(\Omega)} \int_{\Omega} 2\phi^\varepsilon((\phi^\varepsilon)^2 - 1) \, dx.$$

If we set

$$\theta^\varepsilon(x, t) = \varepsilon \lambda^\varepsilon(\phi^\varepsilon(\cdot, t)),$$

(8.13) reduces to

$$\lim_{\varepsilon \downarrow 0} \varepsilon^2 \sup_t |\theta_t^\varepsilon(t)| = 0.$$

We do not know whether this estimate holds. Formally the limiting equation is

$$(11.4) \quad V = \text{mean curvature} + \alpha(t) \quad \text{in } \Omega,$$

with Neumann boundary condition on $\partial\Omega$ (see Giga and Sato [GS]). If Γ_t is a solution of this equation, then

$$\text{Volume enclosed by } \Gamma_t = \int_{\{u(\cdot, t) > 0\}} dx = \frac{1}{2} \lim_{\varepsilon \downarrow 0} \int_{\Omega} [\phi^\varepsilon(x, t) + 1] \, dx.$$

Moreover,

$$\frac{d}{dt} \int_{\Omega} (\phi^\varepsilon + 1) \, dx = \int_{\Omega} \left[\Delta \phi^\varepsilon - \frac{1}{\varepsilon^2} f_0(\phi^\varepsilon) + \lambda^\varepsilon \right] \, dx = 0,$$

i.e., the volume of the region enclosed by Γ_t is constant in time. For a detailed formal analysis of this problem refer to Rubinstein and Sternberg [RS].

The pair $(\Gamma_t, \alpha(t))$ is called a volume preserving mean curvature flow. The associated geometric pde in \mathbb{R}^N is

$$u_t - \left(\Delta u - \frac{(D^2 u Du | Du)}{|Du|^2} \right) - \alpha(t) |Du| = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty).$$

When $N = 2$, Lagrange multiplier $\alpha(t)$ is given by the explicit formula

$$\alpha(t) = 2\pi N(t)/L(t),$$

where $N(t)$ is the number of disjoint of connected parts of $\Gamma(t)$ and $L(t)$ is the length of Γ_t . This formula indicates that the Lagrange multiplier may, in general, be discontinuous in time. If, however, we do not insist that Γ_t is the boundary of a region and replace $\alpha(t)$ by the above formula, then a complete theory is available. In this framework the solution may develop self-intersections, which are not desirable in a physical problem.

In § 4 we presented an example of nonuniqueness for the volume preserving flow by mean curvature.

Example 2 (supercooled Stefan problem). We consider the problem of a melting or growing crystal in a melt. Let $\theta(x, t)$ be the appropriately scaled temperature and $C(t) \subset \mathbb{R}^N$ be the region occupied by the crystal. Gurtin [Gu] derived the equation

$$(11.5) \quad \frac{\partial}{\partial t}[\theta(x, t) + l\mathbb{1}_{C(t)}(x)] = \Delta\theta(x, t) \quad \text{in } (0, \infty) \times \mathbb{R}^N,$$

with the free boundary condition

$$(11.6) \quad \text{normal velocity of } \Gamma_t = \text{curvature} - \theta(x, t) \text{ on } \Gamma_t,$$

where the latent heat $l > 0$ is a given quantity and $\mathbb{1}_{C(t)}$ is the characteristic function of the set $C(t)$. In general, anisotropic versions of the above equation are more appropriate and we refer to Gurtin and Sonner [GuS] for a discussion of the generalizations of (11.5), (11.6), as well as appropriate notion of solution and the underlying physics. Luckhaus [Lu] and Almgren and Wang [AlW] also studied a similar problem in which (11.6) is replaced by the Gibbs–Thompson relation

$$0 = \text{curvature} - \theta \text{ on } \Gamma_t.$$

The system (11.5) and (11.6) can be approximated by the reaction diffusion equations

$$(11.7) \quad \theta_t^\varepsilon + \frac{l}{2} \phi_t^\varepsilon = \Delta\theta^\varepsilon \quad \text{in } \mathbb{R}^N \times (0, \infty),$$

and

$$(11.8) \quad \phi_t^\varepsilon - \Delta\phi^\varepsilon + \frac{1}{\varepsilon^2} \left[f_0(\phi^\varepsilon) - \frac{2}{3} \varepsilon \theta^\varepsilon \right] = 0 \quad \text{in } \mathbb{R}^N \times (0, \infty).$$

The above approximation was first proposed by Caginalp [Ca1]–[Ca3]. The convergence of this system was proved by Caginalp and Chen [CC] in the radial case by a method based on knowing that the limiting motion is classical. Indeed, in the radial case the interface Γ_t is a sphere and (11.6) reduces to an ordinary differential equation. In general, we do not expect Γ_t to be a smooth, classical solution of (11.6). The convergence of the system (11.7)–(11.8) is an open problem.

REFERENCES

- [AAG] S. ALTSHULER, S. ANGENENT, AND Y. GIGA, *Generalized Motion by Mean Curvature for Surfaces of Rotation*, Hokkaido Univ., preprint series no. 119, 1991.
- [A1] S. ANGENENT, *Parabolic equations for curves on surfaces (I): Curves with p -integrable curvature*, Ann. Math., 132 (1990), pp. 171–217.
- [A2] ———, *Parabolic equations for curves on surfaces (II): Intersections, blowup and generalized solutions*, Ann. Math., 133 (1991), pp. 171–217.
- [AlW] F. J. ALMGREN AND L. WANG, *Mathematical Existence of Crystal Growth with Gibbs–Thompson Curvature Effects*, to appear.
- [ArW] D. G. ARONSON AND H. WEINBERGER, *Multidimensional nonlinear diffusion arising in population genetics*, Adv. in Math., 30 (1978), pp. 33–76.
- [Ba1] G. BARLES, *Remark on a Flame Propagation Model*, Rapport 464, 1985.
- [Ba2] ———, *Discontinuous viscosity solutions of first-order Hamilton–Jacobi equations: A guided visit*, Nonlinear Anal. TMA, to appear.
- [BaBs] G. BARLES, L. BRONSARD, AND P. E. SOUGANIDIS, *Front propagation for reaction-diffusion equations of bistable type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.
- [BaES] G. BARLES, L. C. EVANS, AND P. E. SOUGANIDIS, *Wavefront propagation for reaction-diffusion systems of PDE*, Duke Math. J., 61 (1990), pp. 835–859.
- [BaP] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping problems*, Math. Modelling Numer. Anal., 21 (1987), pp. 557–579.
- [BJ1] E. N. BARRON AND R. JENSEN, *Semicontinuous viscosity solutions for Hamilton–Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), 113–174.
- [BJ2] ———, *Optimal Control and Semicontinuous Viscosity Solutions*, preprint, 1990.
- [Br] K. A. BRAKKE, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.
- [BrK] L. BRONSARD AND R. KOHN, *Motion by mean curvature as the singular limit of Ginzburg–Landau model*, J. Differential Equations, 90 (1991), pp. 211–237.
- [Ca1] G. CAGINALP, *An analysis of a phase field model of a free boundary*, Arch. Rational Mech. Anal., 92 (1986), pp. 205–245.
- [Ca2] ———, *Mathematical models of phase boundaries*, in Material Instabilities in Continuum Mechanics and Related Mathematical Problems, J. Ball, ed., Clarendon Press, Oxford, UK, 1988, pp. 35–52.
- [Ca3] ———, *Stefan and Hele–Shaw type models as asymptotic limits of phase field equations*, Phys. Rev. A, 39 (1989), pp. 887–896.
- [CC] G. CAGINALP AND X. CHEN, *Phase Field Equations in the Singular Limit of Sharp Interface Problems*, preprint, 1991.
- [CGG] Y.-G. CHEN, Y. GIGA, AND S. GOTO, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.
- [Ch] X. CHEN, *Generation and propagation of the interface for reaction-diffusion equations*, J. Differential Equations, 96 (1992), pp. 116–141.
- [CEL] M. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
- [CIL] M. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Cahier du CEREMADE na. 9039, 1990; Bull. Amer. Math. Soc., to appear.
- [CL] M. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–43.
- [D] E. DEGEORGI, *Some conjectures on flow by mean curvature*, in Proc. Capri Workshop, 1990.
- [DS] P. DEMOTTONI AND M. SCHATZMAN, *Development of surfaces in \mathbb{R}^N* , Proc. Roy. Soc. Edinburgh Sect. A, 116 (1990), pp. 207–220.
- [ES1] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.
- [ES2] ———, *A PDE approach to geometric optics for certain reaction-diffusion equations*, Indiana Univ. Math. J., 38 (1989), pp. 141–172.
- [ES3] ———, *A PDE approach to certain large deviation problems for systems of parabolic equations*, in Analyse Non Linéaire, Contributions en l'honneur de J.-J. Moreau, Gauthier-Villars, Paris, 1989.
- [ESS] L. C. EVANS, H. M. SONER, AND P. E. SOUGANIDIS, *Phase transitions and generalized motion by curvature*, Comm. Pure Appl. Math., to appear.

- [ESp1] L. C. EVANS AND J. SPRUCK, *Motion of level sets by mean curvature*, J. Differential Geom., 33 (1991), pp. 635–681.
- [ESp2] ———, *Motion of level sets by mean curvature II*, Trans. Amer. Math. Soc., to appear.
- [ESp3] ———, *Motion of level sets by mean curvature III*, J. Geom. Anal., to appear.
- [ESp4] ———, *Motion of level sets by mean curvature IV*, preprint, 1992.
- [Fi] P. C. FIFE, *Dynamics of Internal Layers and Diffusive Interfaces*, CBMS-NSF Regional Conf. Ser. in Appl. Math., No. 53, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988.
- [FiM] P. C. FIFE AND B. MCLEOD, *The approach of solutions of nonlinear differential diffusion equations to travelling solutions*, Arch. Rational Mech. Anal., 65 (1977), pp. 335–361.
- [Fr] M. I. FREIDLIN, *Functional Integration and Partial Differential Equations*, Ann. of Math. Stud., No. 109, Princeton University Press, Princeton, NJ, 1985.
- [Ga] J. GARTNER, *Bistable reaction-diffusion equations and excitable media*, Math. Nachr., 112 (1982), pp. 125–152.
- [GGIS] Y. GIGA, S. GOTO, H. ISHII, AND H. M. SATO, *Comparison principle and convexity preserving properties for singular degenerative parabolic equations on unbounded domains*, Indiana Univ. Math. J., 40 (1991), pp. 443–470.
- [GS] Y. GIGA AND M.-H. SATO, *Generalized interface evolution with the Neumann boundary condition*, Proc. Japan Acad. Ser. A Math. Sci., 67 (1991), pp. 263–266.
- [GT] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, Berlin, 1983.
- [Gu] M. GURTIN, *Multiphase thermomechanics with interfacial structures. I. Heat conduction and the capillary balance law*, Arch. Rational. Mech. Anal., 104 (1988), pp. 195–221.
- [GuS] M. GURTIN AND H. M. SONER, *Some remarks on the Stefan problem with surface structuring*, Quart. Appl. Math., 52 (1992), pp. 291–303.
- [GSS] M. GURTIN, H. M. SONER, AND P. E. SOUGANIDIS, *Aristropic Planar Motion of an Interface Relaxed by the Formation of Infinitesimal Wrinkles*, manuscript.
- [H] R. S. HAMILTON, *Three manifolds with positive Ricci-curvature*, J. Differential Geom., 17 (1982), pp. 255–306.
- [III] T. ILMANEN, *Elliptic Regularization and Partial Regularity for Motion by Mean Curvature*, Ph.D. thesis, Univ. of California, Berkeley, CA, 1991.
- [II2] ———, *Generalized flow of sets by mean curvature on a manifold*, Indiana Univ. Math. J., to appear.
- [Is] H. ISHII, *Hamilton–Jacobi equations with discontinuous Hamiltonians on arbitrary open sets*, Bull. Fac. Sci. Engrg. Chuo Univ., 26 (1985), pp. 5–24.
- [J] R. JENSEN, *The maximum principle for viscosity solutions of second-order fully nonlinear partial differential equations*, Arch. Rational Mech. Anal., 101 (1988), pp. 1–27.
- [JLS] R. JENSEN, P.-L. LIONS, AND P. E. SOUGANIDIS, *A uniqueness result for viscosity solutions of second-order fully nonlinear pde's*, Proc. Amer. Math. Soc., 102 (1988), pp. 975–978.
- [LL] J. M. LASRY AND P.-L. LIONS, *A remark on regularization in Hilbert spaces*, Israel J. Math., 55 (1986), pp. 257–266.
- [Li] P.-L. LIONS, *Generalized Solutions of Hamilton–Jacobi Equations*, Research Notes in Math., No. 69, Pitman, Boston, MA, 1982.
- [Lu] S. LUCKHAUS, *Solutions for the two-phase Stefan problem with Gibbs–Thompson law for the modeling temperature*, European J. Appl. Math., 1 (1990), pp. 101–111.
- [OS] S. Osher AND J. A. SETHIAN, *Fronts moving with curvature-dependent speed: Algorithms based on Hamilton–Jacobi equations*, J. Comput. Phys., 79 (1988), pp. 12–49.
- [RS] J. RUBINSTEIN AND P. STERNBERG, *Nonlocal reaction-diffusion equations and nucleation*, IMA J. Appl. Math., to appear.
- [Se1] J. A. SETHIAN, Ph.D. thesis, Univ. of California, Berkeley, CA, 1984.
- [Se2] ———, *Recent numerical algorithms for hypersurfaces moving with curvature-dependent speed: Hamilton–Jacobi equations and conservation laws*, J. Differential Geom., 31 (1990), pp. 131–162.
- [Se3] ———, *Curvature and evolution of fronts*, Comm. Math. Phys., 101 (1985), pp. 487–499.
- [So] H. M. SONER, *Motion of a set by the mean curvature of its boundary*, J. Differential Equations, to appear.
- [SS] H. M. SONER AND P. E. SOUGANIDIS, *Uniqueness and Singularities of Cylindrically Symmetric Domains Moving by Mean Curvature*, preprint, 1991.
- [Sor] P. Soravia, *Generalized Motion of a Front Along Its Normal Direction: A Differential Games Approach*, preprint, 1992.

EUROPEAN OPTION PRICING WITH TRANSACTION COSTS*

MARK H. A. DAVIS†, VASSILIOS G. PANAS‡, AND THALEIA ZARIPHPOULOU‡

This paper is dedicated to Wendell Fleming on the occasion of his 65th birthday.

Abstract. The authors consider the problem of pricing European options in a market model similar to the Black–Scholes one, except that proportional transaction charges are levied on all sales and purchases of stock. “Perfect replication” is no longer possible, and holding an option involves an essential element of risk. A definition of the option writing price is obtained by comparing the maximum utilities available to the writer by trading in the market with and without the obligation to fulfill the terms of an option contract at the exercise time. This definition reduces to the Black–Scholes value when the transaction costs are removed. Computing the price involves solving two stochastic optimal control problems. This paper shows that the value functions of these problems are the unique viscosity solutions, with different boundary conditions, of a fully nonlinear quasi-variational inequality. This fact implies convergence of discretisation schemes based on the “binomial” approximation of the stock price. Computational results are given. In particular, the authors show that, for a long dated option, the writer must charge a premium over the Black–Scholes price that is just equal to the transaction charge for buying one share.

Key words. option pricing, Black–Scholes formula, transaction costs, utility maximisation, stochastic control, free boundary problem, quasi-variational inequality, viscosity solution, Markov chain approximation

AMS(MOS) subject classifications. primary 35R35, 90A16, 93E20; secondary 35R45, 49B60, 90A09

1. Introduction. Option pricing has been a focus of mathematical research in finance since the publication of the Black–Scholes formula in 1973 [3]. Consult Cox and Rubinstein [5] for a full explanation of financial options and their uses. Black and Scholes gave a valuation for a *European call option*, a contract that confers on the holder the right to buy at the *exercise time* T one share of a specified stock at an agreed price E (known as the *strike price*). Let $S(t)$ denote the stock price at time t . Clearly, the option is worthless if $S(T) \leq E$ but has positive value to the holder, and will be exercised, if $S(T) > E$. The *writer* of the option thus has the obligation to deliver one share at time T for a cash payment of E if $S(T) > E$. The pricing problem is to determine what a buyer should be prepared to pay at some earlier time t to acquire such an option and how much the writer should charge for issuing it.¹ Since holding an option is certainly a speculative position, it seems at first that the answer to this question must depend on the buyer’s or writer’s attitude to risk and therefore that there can be no “universal” pricing formula. However, Black and Scholes showed that, in certain circumstances, such a universal formula is indeed possible. Specifically, they assumed that the stock price process $S(t)$ is a geometric Brownian motion (this is described in § 3, below), that a bank account, i.e., a riskless investment paying interest at a constant rate r , is available, and that funds may be transferred from bank to stock and vice versa without restrictions or costs. Then it turns out that *perfect hedging* is possible: we can form a (time-varying or *dynamic*), self-financing portfolio of holdings in bank and stock whose value at time T is, with probability one, equal to $(S(T) - E)^+$.

* Received by the authors February 18, 1992; accepted for publication (in revised form) May 4, 1992.

† Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BT, England.

‡ Department of Mathematics and Business School, University of Wisconsin, Madison, Wisconsin 53201.

¹ These are in general two separate calculations, although ultimately the writer and the buyer must of course agree on the same price. The circumstances required for such agreement to be reached are not discussed in this paper, but see § 8 for some further remarks.

(This is the value of the option at time T in the absence of any transaction costs, since if $S(T) > E$ the option can be exercised and the stock immediately resold, yielding a profit of $S(T) - E$.) It follows that the value of the option at time $t < T$ is the cash value $w(t)$ of the hedging portfolio at that time. Indeed, suppose that the option is offered for $z < w(t)$; then an investor can proceed as follows. He takes a short position in the hedging portfolio, thus acquiring $w(t)$ at time t , of which z is used to purchase the option, the remainder $(w(t) - z)$ being invested in the bank. At time T the investment in the bank is worth $x := (w(t) - z) \exp(r(T - t))$ and by exercising the option and immediately reselling the stock if $S(T) > E$ the investor acquires $(S(T) - E)^+$. Since the latter is exactly the amount required to close the short position, a sure profit of x has been made. A similar *arbitrage opportunity* is available to the writer if he is able to command a higher price than $w(t)$. It is axiomatic that arbitrage opportunities cannot exist (they contravene any concept of market equilibrium), and therefore $w(t)$ is the unique fair price for the option, from either the buyer's or writer's point of view.

In fact, perfect hedging of very general European contingent claims (ECC) is possible in the Black-Scholes world: if ψ is any function of the stock price trajectory $\{S(u) : t \leq u \leq T\}$ whose expectation exists, then there is a dynamic portfolio whose value at time T is exactly ψ ; this is the *replicating portfolio*. The ability to replicate arbitrary contingent claims is described as *completeness* of the market. In the Black-Scholes world, completeness ultimately hinges on the *martingale representation property* of Brownian motion; see Karatzas [13] for a full explanation. By the same argument as above, the fair price for an ECC with payoff ψ is the initial endowment of the replicating portfolio.

There is a paradoxical element to the Black-Scholes approach, which has been called the "Catch-22 of option pricing theory": the claims that can be priced are just those that are redundant in that the investor could, in principle, simply take a position in the replicating portfolio rather than actually buy the option. Thus, apparently such options have no reason to exist. The fallacy here is that we do not live in a Black-Scholes world. In particular, the replicating portfolio cannot be implemented exactly, since it involves incessant rebalancing, which is impractical in the face of any form of market friction such as transaction costs. In this paper, we develop a theory of option pricing in which transaction costs are explicitly taken into account. Perfect hedging is no longer possible, and therefore *buying or writing options involves an unavoidable element of risk*. For this reason, a preference-independent valuation is no longer possible, and the investor's or writer's attitude toward risk must be considered.

In this paper, a new definition is given for the writing price of a European option, based on utility maximisation theory. This is a modification of the definition introduced by Hodges and Neuberger [11], using a very similar approach. It is also shown that, if a replicating portfolio exists and the class of trading strategies forms a linear space, then the new definition of the writing price coincides with the Black-Scholes price for the contingent claim; in § 3 the Black-Scholes model is stated as an example of a market model where these conditions hold. In § 4 this model is modified to account for transaction costs. We assume for mathematical tractability that investors trade only in the underlying security, although in the presence of transaction costs they might well wish to invest in other securities also. The new definition involves the value functions of two different stochastic control problems and in § 4 the nonlinear partial differential equation (p.d.e.), satisfied by these value functions, is derived using informal arguments. Then we prove that these value functions are the unique viscosity solutions (with the appropriate boundary conditions) of this nonlinear p.d.e. in § 5, and a discretisation scheme is outlined in § 6, together with the proof of convergence to the

unique solution. Finally, § 7 presents the results of this scheme for investors whose preferences are modeled by an exponential utility function (i.e., their index of risk aversion is independent of wealth).

An alternative approach to the pricing problem was outlined by Leland [16], where a hedging strategy is derived based on discrete time rebalancing under transaction costs, but this method is not optimal in any well-defined sense. Also, Bensaid et al. [2] and Edirisinghe, Naik, and Uppal [10] price options using the concept of super-replicating strategies in a discrete time (binomial) model. Our initial reaction is that this approach is unlikely to be viable in a continuous time setting, but this is a question that merits further investigation.

2. Option pricing via utility maximization. In this section, we give a general definition of option price based on utility maximization. This is a modification of work by Hodges and Neuberger [11], who introduced many of the key ideas. The definition can be stated in very general terms, not restricted to any particular market model. The main result of this section, which demonstrates that our approach is well founded, is Theorem 1, which shows that, if perfect replication is possible, then our price coincides with the Black–Scholes price. On the other hand, our definition is applicable in many situations where the Black–Scholes methodology fails.

We consider a time interval $[0, T]$ and a market, which consists of n stocks whose prices $\mathbf{S}(t) = (S_1(t), \dots, S_n(t))$ are assumed to be stochastic processes on a given probability space (Ω, \mathcal{F}, P) ; their natural filtration is $\mathcal{F}_t = \sigma\{S_1(u), \dots, S_n(u) : 0 \leq u \leq t\}$. Investors can also keep their funds in cash, i.e., a risk-free asset, denoted by B . We wish to give a price applicable at time zero for a European option with exercise time T on one of the stocks, say $S_1(t)$.

Let $\mathcal{T}(B)$ denote the set of *admissible trading strategies* available to an investor who starts at time zero with an amount B in cash and no holdings in stock. We identify an element $\pi \in \mathcal{T}(B)$ with a vector stochastic process $(\mathbf{B}^\pi(t), \mathbf{y}^\pi(t)) = (\mathbf{B}^\pi(t), \mathbf{y}_1^\pi(t), \dots, \mathbf{y}_n^\pi(t))$, $t \in [0, T]$, where $\mathbf{B}^\pi(t)$ denotes the amount held in cash and $\mathbf{y}_i^\pi(t)$ the *number of shares* of stock i held, $i = 1, \dots, n$, over $[0, T]$ (this may or may not be constrained to be an integral number). There may be costs associated with transactions. In particular, $c(\mathbf{y}, \mathbf{S})$ is the liquidated cash value of a portfolio vector \mathbf{y} ; i.e., the residual cash value when long positions ($y_i > 0$) are sold, and short positions ($y_i < 0$) closed. We assume that $c(\mathbf{0}, \mathbf{S}) = 0$.

An option on stock $S_1(t)$ is the right to buy one share at time T at a price E , which may be a constant (in the case of a simple call option), or, more generally, may be an \mathcal{F}_T -measurable random variable (allowing for more exotic things such as *look-back* options). We suppose that the option writer forms a portfolio in order to hedge the option and liquidates the portfolio at time T . Suppose that $(B, \mathbf{y}, \mathbf{S})$ denote the cash, stock holdings, and stock price vector at time T , respectively. If $S_1 \leq E$, the option is not exercised, and the cash value of the portfolio is $B + c(\mathbf{y}, \mathbf{S})$. If $S_1 > E$, then the buyer pays the writer E in cash, and the writer delivers one share to the buyer. The value of the writer's portfolio after this transaction is $B + E + c(\mathbf{y} - \mathbf{e}_1, \mathbf{S})$, where \mathbf{e}_1 denotes the vector $(1, 0, \dots, 0)$.

Let $\mathcal{U} : \mathbb{R} \rightarrow \mathbb{R}$, the writer's utility function, be a concave increasing function such that $\mathcal{U}(0) = 0$. It is important that $\mathcal{U}(x)$ is defined for both positive and negative x . Now define the following value function:

$$(2.1) \quad \begin{aligned} V_w(B) = \sup_{\pi \in \mathcal{T}(B)} \mathbb{E} \{ & \mathcal{U}(\mathbf{B}^\pi(T) + I_{(S_1(T) \leq E)} c(\mathbf{y}^\pi(T), \mathbf{S}(T)) \\ & + I_{(S_1(T) > E)} [c(\mathbf{y}^\pi(T) - \mathbf{e}_1, \mathbf{S}(T)) + E]) \}, \end{aligned}$$

where \mathbb{E} denotes expectation, and I_A is the indicator function of the event A . We assume that $V_w(B) < \infty$ for all $B \in \mathbb{R}$ and that $V_w(B)$ is a continuous and monotone-increasing function of B . Note that $V_w(B)$ is the maximum utility available to the writer at time T by liquidating his portfolio after his obligations to the buyer have been met, given an initial endowment B . Now define

$$(2.2) \quad B_w = \inf \{B: V_w(B) \geq 0\}.$$

The writer is thus indifferent between (a) doing nothing and (b) accepting B_w and writing the option. This is, however, not the correct comparison for determining a fair option price. Let

$$(2.3) \quad V_1(B) = \sup_{\pi \in \mathcal{T}(B)} \mathbb{E}\{\mathcal{U}(\mathbf{B}^\pi(T) + c(\underline{\mathbf{y}}^\pi(T), \underline{\mathbf{S}}(T)))\}$$

and define the initial endowment B_1 by

$$(2.4) \quad B_1 = \inf \{B: V_1(B) \geq 0\},$$

where $B_1 \leq 0$, since clearly $V_1(0) \geq 0$. Think of $(-B_1)$ as the “entry fee” that the writer is prepared to pay to get into the market. Our definition of the option writing price p_w is now

$$(2.5) \quad p_w = B_w - B_1.$$

At this price, the writer is indifferent between going into the market to hedge the option and going into the market strictly on his own account. Note that in hedging the option in this way the writer may well hold stocks other than the one on which the option is written.

A primary justification for this definition is that it reduces to the Black-Scholes valuation in cases where this is applicable. Given an option contract on S_1 , a *replicating portfolio* for the writer is an element $\hat{\pi} \in \mathcal{T}(\hat{B})$, for an initial endowment \hat{B} , such that $(\mathbf{B}^{\hat{\pi}}(T), \underline{\mathbf{y}}^{\hat{\pi}}(T)) = I_{(S_1(T) > E)}(-E, e_1)$.

THEOREM 1. *Suppose that the class \mathcal{T} is a linear space; i.e., that if $\pi_i \in \mathcal{T}(B_i)$, $i = 1, 2$ then $a\pi_1 + b\pi_2 \in \mathcal{T}(aB_1 + bB_2)$ for $a, b \in \mathbb{R}$, where $a\pi_1 + b\pi_2 = (a\mathbf{B}^{\pi_1}(t) + b\mathbf{B}^{\pi_2}(t), a\underline{\mathbf{y}}^{\pi_1}(t) + b\underline{\mathbf{y}}^{\pi_2}(t))$. Suppose also that both $V_w(B)$ and $V_1(B)$ are continuous and strictly increasing functions of B . Then $p_w = \hat{B}$ if a replicating portfolio $\hat{\pi} \in \mathcal{T}(\hat{B})$ exists.*

Proof. It follows from the linearity of \mathcal{T} that an arbitrary trading strategy $\pi \in \mathcal{T}(B)$ can always be written in the form $\pi = \hat{\pi} + \tilde{\pi}$, where $\hat{\pi} \in \mathcal{T}(\hat{B})$ and $\tilde{\pi} \in \mathcal{T}(\tilde{B})$ with $B = \hat{B} + \tilde{B}$. Thus, by the continuity and monotonicity assumptions,

$$\begin{aligned} 0 &= V_w(B_w) \\ &= \sup_{\pi \in \mathcal{T}(B_w)} \mathbb{E}\{\mathcal{U}(\mathbf{B}^\pi(T) + I_{(S_1(T) \leq E)}c(\underline{\mathbf{y}}^\pi(T), \underline{\mathbf{S}}(T))) \\ &\quad + I_{(S_1(T) > E)}[E + c(\underline{\mathbf{y}}^\pi(T) - \underline{e}_1, \underline{\mathbf{S}}(T))]\} \\ &= \sup_{\tilde{\pi} \in \mathcal{T}(B_w - \hat{B})} \mathbb{E}\{\mathcal{U}(\mathbf{B}^{\tilde{\pi}}(T) + \mathbf{B}^{\hat{\pi}}(T) + I_{(S_1(T) \leq E)}c(\underline{\mathbf{y}}^{\tilde{\pi}}(T) + \underline{\mathbf{y}}^{\hat{\pi}}(T), \underline{\mathbf{S}}(T))) \\ &\quad + I_{(S_1(T) > E)}[E + c(\underline{\mathbf{y}}^{\tilde{\pi}}(T) + \underline{\mathbf{y}}^{\hat{\pi}}(T) - \underline{e}_1, \underline{\mathbf{S}}(T))]\} \\ &= \sup_{\tilde{\pi} \in \mathcal{T}(B_w - \hat{B})} \mathbb{E}\{\mathcal{U}(\mathbf{B}^{\tilde{\pi}}(T) - EI_{(S_1(T) > E)} + I_{(S_1(T) \leq E)}c(\underline{\mathbf{y}}^{\tilde{\pi}}(T), \underline{\mathbf{S}}(T))) \\ &\quad + I_{(S_1(T) > E)}[E + c(\underline{\mathbf{y}}^{\tilde{\pi}}(T), \underline{\mathbf{S}}(T))]\} \\ &= \sup_{\tilde{\pi} \in \mathcal{T}(B_w - \hat{B})} \mathbb{E}\{\mathcal{U}(\mathbf{B}^{\tilde{\pi}}(T) + c(\underline{\mathbf{y}}^{\tilde{\pi}}(T), \underline{\mathbf{S}}(T)))\} \\ &= V_1(B_w - \hat{B}). \end{aligned}$$

It follows that $B_1 = \inf\{B: V_1(B) \geq 0\}$ is equal to $(B_w - \hat{B})$ and hence that $\hat{B} = B_w - B_1 = p_w$. \square

3. Option pricing without transaction costs. In this section, we show that the standard Black–Scholes model satisfies the conditions of Theorem 1. The stock price and the price of the amount in the bank are described by differential equations, and it is assumed that there are no transaction costs: $c(y, \underline{S}) = y^T \underline{S}$. Several assumptions are made on the admissible trading strategies, which imply that $\mathcal{T}(B)$ is a linear space. For a more detailed presentation of a similar market model, refer to Karatzas [13].

The price of a stock $S_i(t)$, $i = 1, \dots, n$, is modeled by the following Markov diffusion process:

$$(3.1) \quad dS_i(t) = S_i(t)(\alpha_i(t) dt + \sigma_i^T(t) d\mathbf{R}(t)),$$

and the price of the amount in the bank is modeled by the following ordinary differential equation (o.d.e.):

$$(3.2) \quad d\mathbf{B}(t) = \mathbf{r}(t)\mathbf{B}(t) dt,$$

where $\mathbf{R}(t)$ is an n -dimensional P -Brownian motion, which generates the filtration, \mathcal{F}_t , to which $S_i(t)$ is adapted, $\alpha_i(t)$ is the mean growth rate of $S_i(t)$, $\sigma_i^T(t)$ is the i th row of the $n \times n$ volatility matrix σ , and $\mathbf{r}(t)$ is the interest rate. All of these are stochastic processes adapted to \mathcal{F}_t . It is assumed that both $\alpha_i(t)$ and $\mathbf{r}(t)$ are uniformly bounded and that

$$(3.3) \quad \exists \varepsilon > 0 \quad \text{such that } \sigma(t)\sigma^T(t) > \varepsilon I \quad \forall t \in [0, T],$$

where I is the $n \times n$ identity matrix. This last condition is known as the nondegeneracy condition, and it implies that at least one of the n sources of uncertainty in the model affects the price of $S_i(t)$, for each $i = 1, \dots, n$. The set of admissible trading strategies $\mathcal{T}(B)$ consists of the $(n+1)$ -dimensional, right-continuous, measurable, adapted processes, $(\mathbf{B}^\pi(t), \mathbf{y}^\pi(t))$, such that the investor's wealth is bounded below by a nonpositive integrable random variable. The wealth $\mathbf{W}(t)$ obeys the following stochastic differential equation:

$$(3.4) \quad \begin{aligned} d\mathbf{W}(t) &= d\mathbf{B}(t) + \mathbf{y}^T(t) d\mathbf{S}(t) \\ &= \mathbf{r}(t)\mathbf{B}(t) dt + \sum_{i=1}^n \mathbf{y}_i(t) S_i(t) \alpha_i(t) dt + \sum_{i=1}^n \mathbf{y}_i(t) S_i(t) \sigma_i^T(t) d\mathbf{R}(t), \end{aligned}$$

which can also be written as

$$(3.5) \quad d\mathbf{W}(t) = \mathbf{r}(t)\mathbf{W}(t) dt + \sum_{i=1}^n \mathbf{y}_i(t) S_i(t) \sigma_i^T(t) d\tilde{\mathbf{R}}(t),$$

where $\tilde{\mathbf{R}}(t)$ is a Brownian motion with drift.

That $\mathcal{T}(B)$ is a linear space can be verified directly from (3.5). Also, all the value functions are concave and increasing functions of B , as the utility function is a concave and increasing function of the investor's wealth. It can then be easily derived that both value functions are continuous functions of B , and Theorem 1 holds for all the contingent claims, $\psi = (S_1(T) - E)^+$, such that $\mathbb{E}\{\psi^\nu\} < \infty$ for some $\nu > 1$.

Remark. The validity of our price definition may alternatively be proved by deriving expressions for the value functions, following the steps in Karatzas [13].

4. Transaction costs: The Bellman equation for the value functions. It is now assumed that investors must pay transaction costs, which are proportional to the amount

transferred from the stock to the bank. A market model, similar to that of Davis and Norman [8], is then developed, based on the model outlined in the previous section. The main purpose of this section is the derivation of the fully nonlinear p.d.e., actually a variational inequality, satisfied by all the value functions of the utility maximisation problems stated in § 2. Also, a special utility function is defined, the properties of which enable us to determine the dependence of the value functions on the initial endowment B , and thus reduce the dimensionality of the problem.

It is assumed that investors trade only in the underlying stock $S(t)$, on which the contingent claim is written, to further reduce the dimensionality of the problem. Note that, in general, investors may wish to trade in all the risky securities available in the market to maximise their performance criteria. The cash value of a number of shares $y(t)$ of the stock is

$$(4.1) \quad c(y(t), S(t)) = \begin{cases} (1 + \lambda)y(t)S(t), & \text{if } y(t) < 0, \\ (1 - \mu)y(t)S(t), & \text{if } y(t) \geq 0. \end{cases}$$

where λ and μ are the fraction of the traded amount in stock, which the investor pays in transaction costs when buying or selling stock, respectively. The time interval considered is $[0, T]$, and the market model equations are

$$(4.2) \quad d\mathbf{B}(t) = r\mathbf{B}(t) dt - (1 + \lambda)S(t) d\mathbf{L}(t) + (1 - \mu)S(t) d\mathbf{M}(t),$$

$$(4.3) \quad dy(t) = d\mathbf{L}(t) - d\mathbf{M}(t),$$

$$(4.4) \quad dS(t) = S(t)(\alpha dt + \sigma d\mathbf{R}(t)),$$

where $\mathbf{L}(t)$ and $\mathbf{M}(t)$ are the cumulative number of shares bought or sold, respectively, over $[0, T]$ by an investor, $\mathbf{R}(t)$ is a P -Brownian motion that represents the single source of uncertainty, and r , α , and σ are nonrandom constants. As before, \mathcal{F}_t denotes the natural filtration of $\mathbf{R}(t)$. This system of equations describes a degenerate diffusion in \mathbb{R}^3 .

DEFINITION 1. The set of trading strategies $\mathcal{T}(B)$ consists of all the two-dimensional, right-continuous, measurable processes $(\mathbf{B}^\pi(t), y^\pi(t))$, which are the solution of equations (4.2)–(4.4), corresponding to some pair of right-continuous, measurable, \mathcal{F}_t -adapted, increasing processes $(\mathbf{L}(t), \mathbf{M}(t))$, such that

$$(4.5) \quad (\mathbf{B}^\pi(t), y^\pi(t), S(t)) \in \mathcal{E}_K \quad \forall t \in [0, T],$$

where K is a constant, which may depend on the policy π , and

$$(4.6) \quad \mathcal{E}_K = \{(B, y, S) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+: B + c(y, S) > -K\}.$$

By convention, $\mathbf{L}(0-) = \mathbf{M}(0-) = 0$.

Remark. Investors may start with any combination of the two assets at time $s \in [0, T]$ in the general utility maximisation problem, and in that case the class of admissible trading strategies depends on the time s and the initial portfolio, which is characterised by the initial amount in the bank B , an initial number of shares y , and the initial value of the stock S . The constraint (4.5) is required for technical reasons in § 5, and it only rules out strategies that are clearly nonoptimal, as the objective is the maximisation of the utility of final wealth. Also, either $\mathbf{L}(0)$ or $\mathbf{M}(0)$ may be positive; i.e., a jump at the initial time is possible. Finally, (4.2) and (4.3) imply that the trading strategies are self-financing.

Now define the following two functions of wealth at the final time:

$$(4.7) \quad \Phi_1(T, \mathbf{B}(T), y(T), S(T)) = \mathbf{B}(T) + c(y(T), S(T))$$

and

$$(4.8) \quad \Phi_w(T, \mathbf{B}(T), \mathbf{y}(T), \mathbf{S}(T)) = \mathbf{B}(T) + I_{(S_1(T) \leq E)} c(\mathbf{y}(T), \mathbf{S}(T)) \\ + I_{(S_1(T) > E)} [c(\mathbf{y}(T) - 1, \mathbf{S}(T)) + E]$$

and the following value functions:

$$(4.9) \quad V_j(s, B, y, S) = \sup_{\pi \in \mathcal{T}} \mathbb{E}\{\mathcal{U}(\Phi_j(T, \mathbf{B}^\pi(T), \mathbf{y}^\pi(T), \mathbf{S}^\pi(T)))\}$$

where $(s, B, y, S) \in [0, T] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$, and the index j is 1 or w .

From these definitions, it is evident that the dynamic programming algorithm will yield the same p.d.e. for each value function, the terminal condition of which is determined by the utility of the two functions $\Phi_j(T, B, y, S)$, where j is 1 or w . In the following, we derive, at least formally, the Hamilton–Jacobi–Bellman equations, associated with the two stochastic control problems,² which prove to be variational inequalities with gradient constraints. Consider, temporarily, a smaller class of trading strategies \mathcal{T}' , such that $\mathbf{L}(t)$ and $\mathbf{M}(t)$ are absolutely continuous processes, given by

$$(4.10) \quad \mathbf{L}(t) = \int_s^t \mathbf{l}(\xi) d\xi \quad \text{and} \quad \mathbf{M}(t) = \int_s^t \mathbf{m}(\xi) d\xi,$$

where $\mathbf{l}(\xi)$ and $\mathbf{m}(\xi)$ are positive and uniformly bounded by $k < \infty$. Then (4.2)–(4.4) is a vector stochastic differential equation with controlled drift, and the Bellman equation for a value function denoted by V_j^k is

$$(4.11) \quad \max_{0 \leq l, m \leq k} \left\{ \left(\frac{\partial V_j^k}{\partial y} - (1 + \lambda) S \frac{\partial V_j^k}{\partial B} \right) l - \left(\frac{\partial V_j^k}{\partial y} - (1 - \mu) S \frac{\partial V_j^k}{\partial B} \right) m \right\} \\ + \frac{\partial V_j^k}{\partial s} + rB \frac{\partial V_j^k}{\partial B} + \alpha S \frac{\partial V_j^k}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V_j^k}{\partial S^2} = 0$$

for $(s, B, y, S) \in [0, T] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$. The optimal trading strategy is determined by considering the following three possible cases:

$$(4.12) \quad \frac{\partial V_j^k}{\partial y} - (1 + \lambda) S \frac{\partial V_j^k}{\partial B} \geq 0 \quad \text{and} \quad \frac{\partial V_j^k}{\partial y} - (1 - \mu) S \frac{\partial V_j^k}{\partial B} > 0,$$

where the maximum is achieved by $m = 0$ and buying at the maximum possible rate $l = k$;

$$(4.13) \quad \frac{\partial V_j^k}{\partial y} - (1 + \lambda) S \frac{\partial V_j^k}{\partial B} < 0 \quad \text{and} \quad \frac{\partial V_j^k}{\partial y} - (1 - \mu) S \frac{\partial V_j^k}{\partial B} \leq 0,$$

where the maximum is achieved by $l = 0$ and selling at the maximum possible rate $m = k$; and

$$(4.14) \quad \frac{\partial V_j^k}{\partial y} - (1 + \lambda) S \frac{\partial V_j^k}{\partial B} \leq 0 \quad \text{and} \quad \frac{\partial V_j^k}{\partial y} - (1 - \mu) S \frac{\partial V_j^k}{\partial B} \geq 0,$$

where the maximum is achieved by doing nothing; that is $m = 0$ and $l = 0$. (Note that in this case the process $(\mathbf{B}^\pi(t), \mathbf{y}^\pi(t), \mathbf{S}(t))$ becomes an uncontrolled diffusion, which drifts under the influence of the stock process only.) All the other permutations of inequalities are impossible, as all the value functions are increasing functions of B and y .

² The argument is very similar to that in Davis and Norman [8].

The above results suggest that the optimisation problem is a free boundary problem, where, if a value function is known in the four-dimensional space, defined by the state of the investor (s, B, y, S) , the optimal trading strategy is determined by the above inequalities. Also, the state space is divided into three regions, called the *buy*, *sell* and *no-transaction* regions, which are characterised by (4.12), (4.13), and (4.14), respectively. Clearly, the buy and sell regions do not intersect, as it is not optimal to buy and sell at the same time. The boundaries between the no-transaction region and the buy and sell regions are denoted by ∂B and ∂S .

As $k \rightarrow \infty$, the class of admissible trading strategies becomes the class defined at the beginning of this section. It is conjectured that the state space remains divided into a buy region, a sell region, and a no-transaction region, and the optimal trading strategy mandates an immediate transaction to ∂B or ∂S if the state is in the buy region or sell region, followed by transactions of "local time" type at ∂B and ∂S . Therefore, each of the value functions satisfies the following set of equations:

(i) In the buy region, the value function remains constant along the path of the state, dictated by the optimal trading strategy, and therefore

$$(4.15) \quad V_j(s, B, y, S) = V_j(s, B - (1 + \lambda)S \delta y_b, y + \delta y_b, S),$$

where δy_b (the number of shares bought by the investor) can take any positive value up to the number required to take the state to ∂B . Allowing $\delta y_b \rightarrow 0$, (4.15) becomes

$$(4.16) \quad \frac{\partial V_j}{\partial y} - (1 + \lambda)S \frac{\partial V_j}{\partial B} = 0.$$

(ii) Similarly, in the sell region, the value function obeys the following equation:

$$(4.17) \quad V_j(s, B, y, S) = V_j(s, B + (1 - \mu)S \delta y_s, y - \delta y_s, S),$$

where δy_s (the number of shares sold by the investor) can take any positive value up to the number required to take the state to ∂S . Allowing $\delta y_s \rightarrow 0$, (4.17) becomes

$$(4.18) \quad \frac{\partial V_j}{\partial y} - (1 - \mu)S \frac{\partial V_j}{\partial B} = 0.$$

(iii) In the no-transaction region, the value function obeys the same set of equations obtained for the class of absolutely continuous trading strategies, and therefore the value function is given by

$$(4.19) \quad \frac{\partial V_j}{\partial s} + rB \frac{\partial V_j}{\partial B} + \alpha S \frac{\partial V_j}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V_j}{\partial S^2} = 0,$$

and the pair of inequalities, shown above in (4.14), also hold. Note that, due to the continuity of the value function, if it is known in the no-transaction region, it can be determined in both the buy and sell regions by (4.15) and (4.17), respectively.

In the buy region the left-hand side of (4.18) is negative, and, in the sell region, the left-hand side of (4.16) is positive. Also, from the two pairs of inequalities (4.12) and (4.13), it is conjectured that the left-hand side of (4.19) is negative in both the buy and sell regions. Therefore, the above set of equations is condensed into the following fully nonlinear p.d.e.:

$$(4.20) \quad \max \left\{ \frac{\partial V_j}{\partial y} - (1 + \lambda)S \frac{\partial V_j}{\partial B}, - \left(\frac{\partial V_j}{\partial y} - (1 - \mu)S \frac{\partial V_j}{\partial B} \right), \right. \\ \left. \frac{\partial V_j}{\partial s} + rB \frac{\partial V_j}{\partial B} + \alpha S \frac{\partial V_j}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V_j}{\partial S^2} \right\} = 0$$

for $(s, B, y, S) \in [0, T] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$.

Now consider the exponential utility function given by

$$(4.21) \quad \mathcal{U}(x) = 1 - \exp(-\gamma x),$$

where the index of risk aversion is $-\mathcal{U}''(x)/\mathcal{U}'(x) = \gamma$, which is independent of the investor's wealth. The definition of the value functions (4.9) can be written as

$$(4.22) \quad V_j(s, B, y, S) = 1 - \inf_{\pi \in \mathcal{T}} \mathbb{E}\{\exp(-\gamma \mathbf{B}^\pi(T)) \exp(-\gamma \Psi_j(T, \mathbf{y}^\pi(T), \mathbf{S}^\pi(T)))\},$$

where $\Psi_j(T, \mathbf{y}^\pi(T), \mathbf{S}^\pi(T)) := \Phi_j(T, \mathbf{B}^\pi(T), \mathbf{y}^\pi(T), \mathbf{S}^\pi(T)) - \mathbf{B}^\pi(T)$, and the amount $\mathbf{B}^\pi(T)$ is given by the following integral version of the state equation (4.2):

$$(4.23) \quad \mathbf{B}(T) = \frac{B}{\delta(T, s)} - \int_s^T \frac{(1+\lambda)\mathbf{S}(t)}{\delta(T, t)} d\mathbf{L}(t) + \int_s^T \frac{(1-\mu)\mathbf{S}(t)}{\delta(T, t)} d\mathbf{M}(t),$$

where $\delta(T, t)$ is the discount factor, defined by

$$(4.24) \quad \delta(T, t) = \exp(-r(T-t))$$

for the constant interest rate r . Therefore,

$$(4.25) \quad V_j(s, B, y, S) = 1 - \exp\left(-\gamma \frac{B}{\delta(T, s)}\right) Q_j(s, y, S),$$

where $Q_j(s, y, S)$ is a convex nonincreasing continuous function in y and S , defined by $Q_j(s, y, S) = 1 - V_j(s, 0, y, S)$. This representation means that, at time s , the monetary amount invested in the risky asset is independent of the total wealth. It also entails two very important simplifications. First, the writing price is given by the following explicit function of $Q_1(s, 0, S)$ and $Q_w(s, 0, S)$:

$$(4.26) \quad p_w(s, S) = \frac{\delta(T, s)}{\gamma} \ln\left(\frac{Q_w(s, 0, S)}{Q_1(s, 0, S)}\right).$$

Second, (4.20) is transformed into the following p.d.e. for $Q_j(s, y, S)$:

$$(4.27) \quad \min \left\{ \frac{\partial Q_j}{\partial y} + \frac{\gamma(1+\lambda)S}{\delta(T, s)} Q_j, -\left(\frac{\partial Q_j}{\partial y} + \frac{\gamma(1-\mu)S}{\delta(T, s)} Q_j \right), \frac{\partial Q_j}{\partial s} + \alpha S \frac{\partial Q_j}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 Q_j}{\partial S^2} \right\} = 0$$

with boundary conditions, for $Q_1(s, 0, S)$ and $Q_w(s, 0, S)$, given by

$$(4.28) \quad Q_1(T, y, S) = \exp(-\gamma c(y, S))$$

and

$$(4.29) \quad Q_w(T, y, S) = \exp(-\gamma(I_{(S \leq E)} c(y, S) + I_{(S > E)}[c(y-1, S) + E])).$$

Note that the function $Q_j(s, y, S)$ is evaluated in the three-dimensional space $[0, T] \times \mathbb{R} \times \mathbb{R}^+$.

If y defines the vertical axis, the optimal trading strategy when there are no transaction costs is defined by a surface denoted $\mathcal{J}(s, S)$; investors must trade in such a way that $y = \mathcal{J}(s, S)$ at all times. For V_1 , this surface is given by

$$(4.30) \quad \mathcal{J}_1(s, S) = \frac{\delta(T, s)}{\gamma S} \frac{\alpha - r}{\sigma^2}$$

by using the results in Karatzas [13]. For V_w , this surface is given by

$$(4.31) \quad \mathcal{J}_w(s, S) = \frac{\partial \mathcal{C}(s, S)}{\partial S} + \frac{\delta(T, s)}{\gamma S} \frac{\alpha - r}{\sigma^2},$$

where $\mathcal{C}(s, S)$ is the price of the European contingent claim in a market with no transaction costs and the partial derivative with respect to S is the replicating portfolio. In the case with transaction costs, it is conjectured that ∂B and ∂S are two surfaces, $\beta_b(s, S)$ and $\beta_s(s, S)$, which lie strictly below and above $\mathcal{I}(s, S)$, and the no-transaction region is between them (the numerical results, obtained in later sections, support this assertion). If the state of the investor (s, y, S) is in the no-transaction region, then it drifts, under the influence of the diffusion that describes the stock price, on the plane defined by $y = \text{const}$. If the state is in the buy region or sell region, then the investor performs the minimum transaction required to take the state to $\beta_b(s, S)$ or $\beta_s(s, S)$ with an immediate purchase or sale of stock. As noted above, if the function $Q_j(s, y, S)$ is known in the no-transaction region, then (4.16), (4.18), and (4.22) are used to derive

$$(4.32) \quad Q_j(s, y, S) = Q_j(s, y_b, S) \exp\left(-\frac{\gamma(1+\lambda)S}{\delta(T, s)}(y - y_b)\right) \quad \forall y \leq y_b,$$

where $y_b \in \beta_b(s, S)$. A similar equation can be derived for $Q_j(s, y, S)$, for all $y \geq \beta_s(s, S)$.

5. Existence and uniqueness of the solutions of the p.d.e. In this section, we characterise the value functions V_j given by (4.9) as weak (viscosity) solutions of the variational inequality (4.20). Since the stochastic control problems, whose value functions are given by (4.9), are associated with the same Hamilton–Jacobi–Bellman equation, we only examine the problem with value function V_1 . We next show that this value function is a constrained viscosity solution of (4.20) on $[0, T] \times \bar{\mathcal{E}}_\kappa$, where \mathcal{E}_κ is defined by (4.6); the characterisation of V_1 as a constrained viscosity solution of (4.20) is natural due to the presence of the state constraints (4.5).

The notion of *viscosity solutions* was introduced by Crandall and Lions [6] for first-order equations, and by Lions [17] for second-order equations. For a general overview of the theory, we refer to the *user's guide* by Crandall, Ishii, and Lions [7]. Next, we recall the notion of constrained viscosity solutions, which was introduced by Soner [18] and Capuzzo–Dolcetta and Lions [4] for first-order equations (see also Ishii and Lions [12] and Katsoulakis [14]). To this end, we consider a nonlinear second-order p.d.e. of the form

$$(5.1) \quad F(X, W, DW, D^2W) = 0 \quad \text{in } [0, T] \times \mathcal{E},$$

where $\mathcal{E} \subseteq \mathbb{R}^3$, DW , and D^2W denote the gradient vector and the second derivative of W , and the function F is continuous in all its arguments and degenerate elliptic, meaning that

$$(5.2) \quad F(X, p, q, A + N) \leq F(X, p, q, A) \quad \text{if } N \geq 0.$$

DEFINITION 2. A continuous function $W: [0, T] \times \bar{\mathcal{E}} \rightarrow \mathbb{R}$ is a constrained viscosity solution of (5.1) if (i) W is a viscosity subsolution of (5.1) on $[0, T] \times \bar{\mathcal{E}}$; that is, if, for any $\phi \in C^{1,2}([0, T] \times \bar{\mathcal{E}})$ and any local maximum point $X_0 \in [0, T] \times \bar{\mathcal{E}}$ of $W - \phi$,

$$(5.3) \quad F(X_0, W(X_0), D\phi(X_0), D^2\phi(X_0)) \leq 0,$$

and (ii) W is a viscosity supersolution of (5.1) on $[0, T] \times \bar{\mathcal{E}}$; that is, if, for any $\phi \in C^{1,2}([0, T] \times \bar{\mathcal{E}})$ and any local minimum point $X_0 \in [0, T] \times \bar{\mathcal{E}}$ of $W - \phi$,

$$(5.4) \quad F(X_0, W(X_0), D\phi(X_0), D^2\phi(X_0)) \geq 0.$$

THEOREM 2. *The value function $V_1(s, B, y, S)$ is a constrained viscosity solution of*

$$(5.5) \quad \min \left\{ -\left(\frac{\partial W}{\partial y} - (1+\lambda)S \frac{\partial W}{\partial B} \right), \frac{\partial W}{\partial y} - (1-\mu)S \frac{\partial W}{\partial B}, \right. \\ \left. -\left(\frac{\partial W}{\partial s} + rB \frac{\partial W}{\partial B} + \alpha S \frac{\partial W}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 W}{\partial S^2} \right) \right\} = 0$$

on $[0, T] \times \bar{\mathcal{E}}_K$.

Proof. In our case, the state X is (s, x) , where $x = (B, y, S) \in \bar{\mathcal{E}}_K$. Also, (4.20) has been given the above form to turn it into an elliptic p.d.e., to which the uniqueness theorems are applicable. Let $X_0 = (s_0, B_0, y_0, S_0) \in [0, T] \times \bar{\mathcal{E}}_K$: it follows, from results of Zhu [19, Thm. iv.2.2], that there exists an optimal trading strategy, dictated by the pair of processes $(\mathbf{L}^*(t), \mathbf{M}^*(t))$, where $\mathbf{X}_0^*(t) = (t, \mathbf{B}_0^*(t), \mathbf{y}_0^*(t), \mathbf{S}_0^*(t))$ is the optimal trajectory, with $\mathbf{X}_0^*(s_0) = X_0$.

(i) First, we prove that V_1 is a viscosity subsolution of (5.5) on $[0, T] \times \bar{\mathcal{E}}_K$; for this, we must show that, for all smooth functions $\phi(X)$, such that $V_1(X) - \phi(X)$ has a local maximum at $X_0 \in [0, T] \times \bar{\mathcal{E}}_K$, the following inequality holds:

$$(5.6) \quad \min \left\{ -\left(\frac{\partial \phi(X_0)}{\partial y} - (1+\lambda)S_0 \frac{\partial \phi(X_0)}{\partial B} \right), \frac{\partial \phi(X_0)}{\partial y} - (1-\mu)S_0 \frac{\partial \phi(X_0)}{\partial B}, \right. \\ \left. -\left(\frac{\partial \phi(X_0)}{\partial s} + rB_0 \frac{\partial \phi(X_0)}{\partial B} + \alpha S_0 \frac{\partial \phi(X_0)}{\partial S} + \frac{1}{2} \sigma^2 S_0^2 \frac{\partial^2 \phi(X_0)}{\partial S^2} \right) \right\} \leq 0.$$

Without loss of generality, we assume that $V_1(X_0) = \phi(X_0)$ and $V_1 \leq \phi$ on $[0, T] \times \bar{\mathcal{E}}_K$. We argue by contradiction: if the arguments inside the minimum operator of (5.6) satisfy

$$(5.7) \quad \frac{\partial \phi(X_0)}{\partial y} - (1+\lambda)S_0 \frac{\partial \phi(X_0)}{\partial B} < 0,$$

$$(5.8) \quad \frac{\partial \phi(X_0)}{\partial y} - (1-\mu)S_0 \frac{\partial \phi(X_0)}{\partial B} > 0,$$

then there exists $\theta > 0$, such that

$$(5.9) \quad \frac{\partial \phi(X_0)}{\partial s} + rB_0 \frac{\partial \phi(X_0)}{\partial B} + \alpha S_0 \frac{\partial \phi(X_0)}{\partial S} + \frac{1}{2} \sigma^2 S_0^2 \frac{\partial^2 \phi(X_0)}{\partial S^2} < -\theta.$$

From the fact that ϕ is smooth, the above inequalities become

$$(5.10) \quad \frac{\partial \phi(X)}{\partial y} - (1+\lambda)S \frac{\partial \phi(X)}{\partial B} < 0,$$

$$(5.11) \quad \frac{\partial \phi(X)}{\partial y} - (1-\mu)S \frac{\partial \phi(X)}{\partial B} > 0,$$

and

$$(5.12) \quad \frac{\partial \phi(X)}{\partial s} + rB \frac{\partial \phi(X)}{\partial B} + \alpha S \frac{\partial \phi(X)}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 \phi(X)}{\partial S^2} < -\theta,$$

where $X = (s, B, y, S) \in \mathcal{B}(X_0)$, a neighborhood of X_0 . In Lemma 1 it is shown that $\mathbf{X}_0^*(t)$ has no jumps, P -a.s., at $X_0 = \mathbf{X}_0^*(s_0)$. Hence, $\tau(\omega)$, defined by

$$(5.13) \quad \tau(\omega) = \inf \{t \in [s_0, T]: \mathbf{X}_0^*(t) \notin \mathcal{B}(X_0)\},$$

is positive P -a.s., and therefore

$$\begin{aligned}
 -\theta \mathbb{E}\{\tau\} &\geq \mathbb{E} \int_{s_0}^{\tau} \left(\frac{\partial \phi(\mathbf{X}_0^*(t))}{\partial y} - (1+\lambda) \mathbf{S}_0^*(t) \frac{\partial \phi(\mathbf{X}_0^*(t))}{\partial B} \right) d\mathbf{L}^*(t) \\
 &\quad - \mathbb{E} \int_{s_0}^{\tau} \left(\frac{\partial \phi(\mathbf{X}_0^*(t))}{\partial y} - (1-\mu) \mathbf{S}_0^*(t) \frac{\partial \phi(\mathbf{X}_0^*(t))}{\partial B} \right) d\mathbf{M}^*(t) \\
 (5.14) \quad &\quad + \mathbb{E} \int_{s_0}^{\tau} \left(\frac{\partial \phi(\mathbf{X}_0^*(t))}{\partial s} + r \mathbf{B}_0^*(t) \frac{\partial \phi(\mathbf{X}_0^*(t))}{\partial B} \right. \\
 &\quad \left. + \alpha \mathbf{S}_0^*(t) \frac{\partial \phi(\mathbf{X}_0^*(t))}{\partial S} + \frac{1}{2} \sigma^2 (\mathbf{S}_0^*(t))^2 \frac{\partial^2 \phi(\mathbf{X}_0^*(t))}{\partial S^2} \right) dt \\
 &= \mathbb{E}\{I_1\} - \mathbb{E}\{I_2\} + \mathbb{E}\{I_3\},
 \end{aligned}$$

where $(\mathbf{L}^*(t), \mathbf{M}^*(t))$ is the optimal trading strategy at X_0 . Applying Itô's formula to $\phi(X)$, where the state dynamics are given by (4.2)–(4.4), we get

$$(5.15) \quad \mathbb{E}\{\phi(\mathbf{X}_0^*(\tau))\} = \phi(X_0) + \mathbb{E}\{I_1\} - \mathbb{E}\{I_2\} + \mathbb{E}\{I_3\}.$$

Since $V(X) \leq \phi(X)$, for all $X \in \mathcal{B}(X_0)$, and $V_1(X_0) = \phi(X_0)$, (5.14) and (5.15) yield

$$(5.16) \quad \mathbb{E}\{V_1(\mathbf{X}_0^*(\tau))\} \leq V_1(X_0) + (\mathbb{E}\{I_1\} - \mathbb{E}\{I_2\} + \mathbb{E}\{I_3\}) < V_1(X_0) - \theta \mathbb{E}\{\tau\},$$

which violates the dynamic programming principle, together with the optimality of $(\mathbf{L}^*(t), \mathbf{M}^*(t))$. Therefore, at least one of the arguments inside the minimum operator of (5.6) is nonpositive, and hence the value function is a viscosity subsolution of (5.5).

(ii) In the second part of the proof, we show that V_1 is a viscosity supersolution of (5.5) in $[0, T] \times \mathcal{E}_K$; for this we must show that, for all smooth functions $\phi(X)$, such that $V_1(X) - \phi(X)$ has a local minimum at $X_0 \in [0, T] \times \mathcal{E}_K$, the following inequality holds:

$$\begin{aligned}
 (5.17) \quad \min \left\{ - \left(\frac{\partial \phi(X_0)}{\partial y} - (1+\lambda) S_0 \frac{\partial \phi(X_0)}{\partial B} \right), \frac{\partial \phi(X_0)}{\partial y} - (1-\mu) S_0 \frac{\partial \phi(X_0)}{\partial B}, \right. \\
 \left. - \left(\frac{\partial \phi(X_0)}{\partial s} + r B_0 \frac{\partial \phi(X_0)}{\partial B} + \alpha S_0 \frac{\partial \phi(X_0)}{\partial S} + \frac{1}{2} \sigma^2 S_0^2 \frac{\partial^2 \phi(X_0)}{\partial S^2} \right) \right\} \geq 0,
 \end{aligned}$$

where, without loss of generality, $V_1(X_0) = \phi(X_0)$ and $V_1 \geq \phi$ on $[0, T] \times \bar{\mathcal{E}}_K$. In this case, we prove that each argument of the minimum operator of (5.17) is nonnegative.

Consider the trading strategy $\mathbf{L}(t) = L_0 > 0$, $s_0 \leq t \leq T$, and $\mathbf{M}(t) = 0$, $s_0 \leq t \leq T$. By the dynamic programming principle,

$$(5.18) \quad V_1(s_0, B_0, y_0, S_0) \geq V_1(s_0, B_0 - (1+\lambda) S_0 L_0, y_0 + L_0, S_0).$$

This inequality holds for $\phi(s, B, y, S)$ as well, and, by taking the left-hand side to the right-hand side, dividing by L_0 , and sending $L_0 \rightarrow 0$, we get

$$(5.19) \quad \frac{\partial \phi(X_0)}{\partial y} - (1+\lambda) S_0 \frac{\partial \phi(X_0)}{\partial B} \geq 0.$$

Similarly, by using the trading strategy $\mathbf{L}(t) = 0$, $s_0 \leq t \leq T$, and $\mathbf{M}(t) = M_0 > 0$, $s_0 \leq t \leq T$, the second argument inside the minimum operator is found to be nonnegative.

Finally, consider the case where no trading is applied. By the dynamic programming principle

$$(5.20) \quad \mathbb{E}\{V_1(\mathbf{X}_0^d(t))\} \leq V_1(s_0, B_0, y_0, S_0),$$

where $\mathbf{X}_0^d(t)$ is the state trajectory when $\mathbf{M}(t) = \mathbf{L}(t) = 0$, $s_0 \leq t \leq T$, given by (4.2)–(4.4) as

$$(5.21) \quad \mathbf{X}_0^d(t) = (t, B_0 \exp(r(t-s_0)), y_0, S_0 \exp((\alpha - \frac{1}{2}\sigma^2)(t-s_0) + \sigma(\mathbf{R}(t) - \mathbf{R}(s_0))))$$

and $\mathbf{X}_0^d(t) \in \mathcal{B}(X_0)$. Therefore, by applying Itô's rule on $\phi(s, B, y, S)$, inequality (5.20) yields

$$(5.22) \quad \mathbb{E} \left\{ \int_{s_0}^t \left(\frac{\partial \phi(\mathbf{X}_0^d(\xi))}{\partial s} + r \mathbf{B}_0^d(\xi) \frac{\partial \phi(\mathbf{X}_0^d(\xi))}{\partial B} + \alpha \mathbf{S}_0^d(\xi) \frac{\partial \phi(\mathbf{X}_0^d(\xi))}{\partial S} + \frac{1}{2} \sigma^2 (\mathbf{S}_0^d(\xi))^2 \frac{\partial^2 \phi(\mathbf{X}_0^d(\xi))}{\partial S^2} \right) d\xi \right\} \geq 0,$$

and, by sending $t \rightarrow s_0$, the third argument inside the minimum operator is found to be nonnegative (for detailed proof, see Lions [17]). This completes the proof. \square

LEMMA 1. Assume that inequality (5.7) holds and denote the event that the optimal trajectory $\mathbf{X}_0^*(t)$ has a jump of size ε , along the direction $(0, -(1+\lambda)S_0, 1, 0)$, by $A(\omega)$. Assume that the state (after the jump) is $(s_0, B_0 - (1+\lambda)S_0\varepsilon, y_0 + \varepsilon, S_0) \in \mathcal{B}(X_0)$. Then

$$(5.23) \quad \left(\frac{\partial \phi(X_0)}{\partial y} - (1+\lambda)S \frac{\partial \phi(X_0)}{\partial B} \right) P(A) \geq 0,$$

and therefore $P(A) = 0$. Similarly, if inequality (5.8) holds, then the optimal trajectory has no jumps along the direction $(0, (1-\mu)S_0, -1, 0)$, P -a.s. at X_0 .

Proof. By the principle of dynamic programming,

$$(5.24) \quad \begin{aligned} V_1(s_0, B_0, y_0, S_0) &= \mathbb{E}\{V_1(s_0, B_0 - (1+\lambda)S_0\varepsilon, y_0 + \varepsilon, S_0)\} \\ &= \int_{A(\omega)} V_1(s_0, B_0 - (1+\lambda)S_0\varepsilon, y_0 + \varepsilon, S_0) dP \\ &\quad + \int_{\Omega - A(\omega)} V_1(s_0, B_0, y_0, S_0) dP, \end{aligned}$$

and therefore

$$(5.25) \quad \int_{A(\omega)} (\phi(s_0, B_0 - (1+\lambda)S_0\varepsilon, y_0 + \varepsilon, S_0) - \phi(s_0, B_0, y_0, S_0)) dP \geq 0,$$

since $V_1(X) \leq \phi(X)$ for all $X \in \mathcal{B}(X_0)$ and $V_1(X_0) = \phi(X_0)$. Therefore,

$$(5.26) \quad \limsup_{\varepsilon \rightarrow 0} \left\{ \int_{A(\omega)} \frac{\phi(s_0, B_0 - (1+\lambda)S_0\varepsilon, y_0 + \varepsilon, S_0) - \phi(s_0, B_0, y_0, S_0)}{\varepsilon} dP \right\} \geq 0,$$

and, by Fatou's lemma,

$$(5.27) \quad \int_{A(\omega)} \limsup_{\varepsilon \rightarrow 0} \left\{ \frac{\phi(s_0, B_0 - (1+\lambda)S_0\varepsilon, y_0 + \varepsilon, S_0) - \phi(s_0, B_0, y_0, S_0)}{\varepsilon} \right\} dP \geq 0,$$

which implies (5.23). \square

This section is concluded by showing that the value function V_1 is the unique bounded constrained viscosity solution of (5.5). Since this uniqueness result will be used mainly for the convergence of the numerical scheme presented in the next section, we prove the theorem for the exponential utility function. For simplicity of exposition, we assume that the interest rate $r = 0$. The argument is the same but notationally more cumbersome when $r > 0$.

THEOREM 3. *Let u be a bounded upper semicontinuous viscosity subsolution of (5.5) on $[0, T] \times \bar{\mathcal{E}}_K$, and let v be a bounded from below lower semicontinuous viscosity supersolution of (5.5) in $[0, T] \times \mathcal{E}_K$, such that $u(T, x) \leq v(T, x)$, for all $x \in \bar{\mathcal{E}}_K$, and $u(t, B, y, 0) \leq v(t, B, y, 0)$, on $[0, T] \times \bar{\mathcal{E}}_K$, where $u(T, x) = 1 - \exp(-\gamma(B + c(y, S)))$ and $u(t, B, y, 0) = 1 - \exp(-\gamma B)$. Then $u \leq v$ on $[0, T] \times \bar{\mathcal{E}}_K$.*

Note. The proof relies on arguments used in § v of Ishii and Lions [12]; only the main steps are presented.

Proof. Sketch. We first construct a positive strict supersolution of (5.5) in $[0, T] \times \mathcal{E}_K$. To this end, let $h: [0, T] \times \mathcal{E}_K \rightarrow \mathbb{R}^+$ be given by $h(t, B, y, S) = 1 - \exp(-\gamma(B + kyS)) + C_1(T - t) + C_2$, where the constants k , C_1 , and C_2 satisfy

$$(5.28) \quad 1 + \lambda > k > 1 - \mu, \quad C_1 > \frac{\alpha^2}{2\sigma^2} \exp(\gamma K), \quad \text{and } C_2 > \exp(K) - 1.$$

Then

$$(5.29) \quad \begin{aligned} & H(X, h_t, Dh, D^2h) \\ &= \min \left\{ -\frac{\partial h}{\partial y} + (1 + \lambda)S \frac{\partial h}{\partial B}, \frac{\partial h}{\partial y} - (1 - \mu)S \frac{\partial h}{\partial B}, -\frac{\partial h}{\partial s} - \alpha S \frac{\partial h}{\partial S} - \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 h}{\partial S^2} \right\} \\ &= \exp(-\gamma(B + kyS)) \\ &\quad \cdot \min \left\{ \gamma S(1 + \lambda - k), \gamma S(k - (1 - \mu)), \right. \\ &\quad \left. C_1 \exp(\gamma(B + kyS)) + \frac{1}{2} \gamma^2 k^2 \sigma^2 (yS)^2 - \alpha k \gamma (yS) \right\}. \end{aligned}$$

Using (5.28) and the fact that the minimum value of the quadratic $\mathcal{Q}(\xi) = \frac{1}{2} \gamma^2 k^2 \sigma^2 \xi^2 - \alpha \gamma k \xi$ is $-\alpha^2/2\sigma^2$, the above inequality yields

$$(5.30) \quad \begin{aligned} & H(X, h_t, Dh, D^2h) \\ &> \exp(-\gamma(B + kyS)) \min \{ \gamma S(1 + \lambda - k), \gamma S(k - (1 - \mu)), K' \} \end{aligned}$$

in $[0, T] \times \bar{\mathcal{E}}_K$, where $0 < K' < C_1 - (\alpha^2/2\sigma^2) \exp(\gamma K)$.

Therefore, h is a strict supersolution (5.5). The fact that $h > 0$ follows from the choice of the constant C_2 .

To conclude the proof of the theorem, we will need the following key lemma. Its proof follows along the lines of Theorem vi.5 in Ishii and Lions [12], and therefore it is omitted.

LEMMA 2. *Let u be a bounded lower semicontinuous viscosity subsolution of (5.5) on $[0, T] \times \bar{\mathcal{E}}_K$, and let v be a bounded from below uniformly continuous viscosity supersolution of (5.5) in $[0, T] \times \mathcal{E}_K$ of the equation $H(X, v_t, Dv, D^2v) - f(x) = 0$, where $f > 0$ in \mathcal{E}_K , $u(T, x) \leq v(T, x)$, for all $x \in \bar{\mathcal{E}}_K$, and $u(t, B, y, 0) \leq v(t, B, y, 0)$, on $[0, T] \times \bar{\mathcal{E}}_K$. Then $u \leq v$ on $[0, T] \times \bar{\mathcal{E}}_K$.*

We now conclude the proof of the theorem. To this end, we first observe that, because of the choice of k , C_1 and C_2 ,

$$(5.31) \quad h(T, B, y, S) > 1 - \exp(-\gamma(B + c(y, S))) \quad \text{and} \quad h(t, B, y, 0) > 1 - \exp(-\gamma B).$$

Next, we define the function $w^\theta = \theta v + (1 - \theta)h$, where $0 < \theta < 1$, and, using (5.31), we get

$$(5.32) \quad w^\theta(T, B, y, S) \geq u(T, B, y, S) \quad \text{and} \quad w^\theta(t, B, y, 0) \geq u(t, B, y, 0).$$

We also observe that w^θ is a viscosity supersolution of $H - g = 0$, in $[0, T] \times \bar{\mathcal{E}}_K$, where $g = (1 - \theta)f$. In fact, let $\psi \in C^{1,2}([0, T] \times \mathcal{E}_K)$ and assume that $w^\theta - \psi$ has a minimum at X_0 . Then $v - \phi$ also has a minimum at X_0 , where $\phi = (1/\theta) \times (\psi - (1 - \theta)h)$. Using the fact that v is a viscosity supersolution of $H = 0$ and (5.30) we get

$$(5.33) \quad \begin{aligned} &\theta H(X_0, \phi_t(X_0), D\phi(X_0), D^2\phi(X_0)) \\ &+ (1 - \theta)H(X_0, h_t(X_0), Dh(X_0), D^2h(X_0)) \geq (1 - \theta)f(X_0). \end{aligned}$$

As the Hamiltonian $H(x, p, q, A)$ is jointly concave with respect to (p, q, A) , (5.33) yields

$$(5.34) \quad H(X_0, \psi_t(X_0), D\psi(X_0), D^2\psi(X_0)) \geq (1 - \theta)f(X_0),$$

which in turn implies that w^θ is a viscosity supersolution of $H - g = 0$. Finally, applying Lemma 2 to u and w^θ , we get

$$(5.35) \quad u \leq w^\theta \quad \text{on } [0, T] \times \bar{\mathcal{E}}_K,$$

and sending $\theta \uparrow 1$ concludes the proof of the theorem. \square

6. Discretisation and solution of the problem. The solution of the p.d.e. (4.20) is obtained by turning the stochastic differential equations (4.2)–(4.4) into Markov chains to apply the discrete time dynamic programming algorithm. The method here closely follows the one given in Martins and Kushner [15]. The discrete state is $\mathbb{X} = (\iota, \mathbb{B}, \eta, \mathbb{S})$, whose elements denote time, amount in the bank, number of shares, and stock price in a discrete space. The value functions, denoted by \mathbb{V}_1 and \mathbb{V}_w , are given a value at the final time by using the boundary conditions for the continuous value functions over the discrete subspace $(\mathbb{B}, \eta, \mathbb{S})$, and then they are estimated by proceeding backward in time by using the discrete time algorithm. As in the continuous time case, this algorithm is the same for both value functions and is derived below for a value function denoted by $\mathbb{V}_j^\rho(\iota, \mathbb{B}, \eta, \mathbb{S})$, where ρ is a discretisation parameter, which depends on the discrete time interval δt . If δt and the resolution of the η axis $\delta\eta$ are sent to zero, then the above discrete value function converges to a viscosity subsolution and a viscosity supersolution of the p.d.e. (4.20). Therefore, all the discrete value functions converge to their continuous counterparts; this is due to the uniqueness of the viscosity solution.

The discrete time variable ι takes values in $\{0, \delta t, 2\delta t, \dots, N\delta t\}$, where δt is the discrete time interval and $T - s = N\delta t$. The Markov chain for the discrete stock price process $\mathbb{S}(\iota)$ is modeled by

$$(6.1) \quad \mathbb{S}(\iota + 1) = \begin{cases} \mathbb{S}(\iota) \times k_u & \text{with probability } \frac{1}{2}, \\ \mathbb{S}(\iota) \times k_d & \text{with probability } \frac{1}{2}, \end{cases}$$

where the values of k_u and k_d are determined by equating the first and second moments of the chain with those of the diffusion, which describes S , and therefore

$$(6.2) \quad k_u = \exp(\alpha \delta t + \sigma \sqrt{\delta t}) \quad \text{and} \quad k_d = \exp(\alpha \delta t - \sigma \sqrt{\delta t}).$$

The discretisation scheme and its convergence properties are more thoroughly explained in Chapter 5 of Cox and Rubinstein [5]. The discrete time equation for the amount in the bank $\mathbb{B}(\iota)$ is

$$(6.3) \quad \mathbb{B}(\iota + 1) = \mathbb{B}(\iota) \exp(r \delta t),$$

which is a deterministic difference equation.

The discrete time dynamic programming principle is invoked, and the following discretisation scheme is proposed for the p.d.e. (4.20):

$$(6.4) \quad \mathcal{J}(\rho) \mathbb{V}_j^\rho - \mathbb{V}_j^\rho = 0,$$

where $\mathcal{S}(\rho)$ is an operator given by

$$(6.5) \quad \mathcal{S}(\rho)\mathbb{V}_j^\rho = \max \{ \mathbb{V}_j^\rho(\iota, \mathbb{B} - (1 + \lambda)\mathbb{S}\kappa\rho, \eta + \kappa\rho, \mathbb{S}), \mathbb{V}_j^\rho(\iota, \mathbb{B} + (1 - \mu)\mathbb{S}\kappa\rho, \eta - \kappa\rho, \mathbb{S}), \\ \mathbb{E}\{ \mathbb{V}_j^\rho(\iota + \rho, \mathbb{B} \exp(r\rho), \eta, \mathbb{S} \exp(\alpha\rho + \theta\sigma\sqrt{\rho})) \} \},$$

where $\rho = \delta t$, κ is a real constant and θ is a random variable taking values ± 1 with probability $\frac{1}{2}$ each. This scheme is based on the principle that the investor's policy is the choice of the optimum transaction, that is, to buy or sell or do nothing for a particular state given the value function for all the states in the next time instant. We next show that, as the discretisation parameter $\rho \rightarrow 0$, the solution \mathbb{V}_j^ρ of (6.4) converges to the value function V , or, equivalently, to the unique constrained viscosity solution of (4.20). Although the proof of the theorem follows along the lines of Barles and Souganidis [Thm. 2.1, p. 1], it is presented here for completeness.

THEOREM 4. *The solution \mathbb{V}_j^ρ of (6.4) converges locally uniformly as $\rho \rightarrow 0$ to the unique continuous constrained viscosity solution of (4.20).*

Proof. Let

$$(6.6) \quad V_j^\rho(t, B, y, S) = \begin{cases} \mathbb{V}_j^\rho(\iota, \mathbb{S}, \eta, \mathbb{B}), & \text{if } t \in [\iota, \iota + \rho), y \in [\eta, \eta + \kappa\rho), \\ \Phi_j(\mathbb{B}, \eta, \mathbb{S}), & \text{if } t = T \end{cases},$$

and

$$(6.7) \quad \underline{V}_j(X) = \liminf_{Y \rightarrow X, \rho \rightarrow 0} \{ \mathbb{V}_j^\rho(Y) \} \quad \text{and} \quad \bar{V}_j(X) = \limsup_{Y \rightarrow X, \rho \rightarrow 0} \{ \mathbb{V}_j^\rho(Y) \},$$

where $X = (t, B, y, S)$. We will show that \underline{V}_j and \bar{V}_j are a viscosity supersolution and a viscosity subsolution of (4.20), respectively. Combining this with the uniqueness result of Theorem 3 yields $\underline{V}_j \geq \bar{V}_j$ on $[0, T] \times \mathcal{E}_K$. The opposite inequality is true by the definition of \underline{V}_j and \bar{V}_j , and therefore,

$$(6.8) \quad \underline{V}_j(X) = \bar{V}_j(X) = V_j(X),$$

which, together with (6.7), also implies the local uniform convergence of \mathbb{V}_j^ρ to V_j (see [1]).

We will only prove that \underline{V}_j is a viscosity supersolution of (4.20), since the arguments for \bar{V}_j are identical. Let X_0 be a local minimum of $\underline{V}_j - \phi$ on $[0, T] \times \mathcal{E}_K$, for $\phi \in C^{1,2}([0, T] \times \mathcal{E}_K)$. Without loss of generality, we may assume that X_0 is a strict local minimum, that $\underline{V}_j(X_0) = \phi(X_0)$, and that $\phi \leq -2 \times \sup_\rho \{ \|\mathbb{V}_j^\rho\|_\infty \}$ outside the ball $\mathcal{B}(X_0, R)$, $R > 0$, where $\underline{V}_j(X) - \phi(X) \geq 0$.

Then there exist sequences $\rho_n \in \mathbb{R}^+$ and $Y_n \in [0, T] \times \mathcal{E}_K$, such that

$$(6.9) \quad \rho_n \rightarrow 0, Y_n \rightarrow X_0, \mathbb{V}_j^{\rho_n}(Y_n) \rightarrow \underline{V}_j(X_0), Y_n \text{ is a global minimum point of } \mathbb{V}_j^{\rho_n} - \phi.$$

Let $h_n = \mathbb{V}_j^{\rho_n} - \phi$; then

$$(6.10) \quad h_n \rightarrow 0 \quad \text{and} \quad \mathbb{V}_j^{\rho_n}(X) \geq \phi(X) + h_n(X) \quad \forall X \in [0, T] \times \mathcal{E}_K.$$

To show that \underline{V}_j is a viscosity supersolution of (4.20), it suffices to show that

$$(6.11) \quad \min \left\{ - \left(\frac{\partial \phi(X_0)}{\partial y} - (1 + \lambda) S_0 \frac{\partial \phi(X_0)}{\partial B} \right), \left(\frac{\partial \phi(X_0)}{\partial y} - (1 - \mu) S_0 \frac{\partial \phi(X_0)}{\partial B} \right), \right. \\ \left. - \left(\frac{\partial \phi(X_0)}{\partial s} + r B_0 \frac{\partial \phi(X_0)}{\partial B} + \alpha S_0 \frac{\partial \phi(X_0)}{\partial S} + \frac{1}{2} \sigma^2 S_0^2 \frac{\partial^2 \phi(X_0)}{\partial S^2} \right) \right\} \geq 0.$$

Let $Y_n = (t_{\rho_n}, \mathbb{B}_{\rho_n}, y_{\rho_n}, \mathbb{S}_{\rho_n})$, where $t_{\rho_n} \in [\iota_{\rho_n}, \iota_{\rho_n} + \rho_n)$ and $y_{\rho_n} \in [\eta_{\rho_n}, \eta_{\rho_n} + \kappa\rho_n)$. Then

$$\begin{aligned}
 & \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, y_{\rho_n}, \mathbb{S}_{\rho_n}) \\
 &= \max \{ \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n} - (1+\lambda)\mathbb{S}_{\rho_n}\kappa\rho_n, \eta_{\rho_n} + \kappa\rho_n, \mathbb{S}_{\rho_n}), \\
 (6.12) \quad & \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n} + (1-\mu)\mathbb{S}_{\rho_n}\kappa\rho_n, \eta_{\rho_n} - \kappa\rho_n, \mathbb{S}_{\rho_n}), \\
 & \mathbb{E}\{\mathbb{V}_j^{\rho_n}(\iota_{\rho_n} + \rho_n, \mathbb{B}_{\rho_n} \exp(r\rho_n), \eta_{\rho_n}, \mathbb{S}_{\rho_n} \exp(\alpha\rho_n + \theta\sigma\sqrt{\rho_n}))\} \}.
 \end{aligned}$$

Now we look at the following three cases.

Case 1. It holds that

$$\mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}) = \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n} - (1+\lambda)\mathbb{S}_{\rho_n}\kappa\rho_n, \eta_{\rho_n} + \kappa\rho_n, \mathbb{S}_{\rho_n}).$$

Then (6.10) implies that

$$\begin{aligned}
 (6.13) \quad & \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}) \geq \phi(\iota_{\rho_n}, \mathbb{B}_{\rho_n} - (1+\lambda)\mathbb{S}_{\rho_n}\kappa\rho_n, \eta_{\rho_n} + \kappa\rho_n, \mathbb{S}_{\rho_n}) \\
 & + \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}) - \phi(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}),
 \end{aligned}$$

and therefore

$$\begin{aligned}
 0 & \geq \liminf_n \left\{ \frac{\phi(\iota_{\rho_n}, \mathbb{B}_{\rho_n} - (1+\lambda)\mathbb{S}_{\rho_n}\kappa\rho_n, \eta_{\rho_n} + \kappa\rho_n, \mathbb{S}_{\rho_n}) - \phi(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n})}{\rho_n} \right\} \\
 (6.14) \quad & \geq \liminf_{\rho \rightarrow 0} \left\{ \frac{\phi(t_0, B_0 - (1+\lambda)S_0\kappa\rho, y_0 + \kappa\rho, S_0) - \phi(t_0, B_0, y_0, S_0)}{\rho} \right\} \\
 & = \frac{\partial \phi(X_0)}{\partial y} - (1+\lambda)S_0 \frac{\partial \phi(X_0)}{\partial B}.
 \end{aligned}$$

Case 2. It holds that

$$\mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}) = \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n} + (1-\mu)\mathbb{S}_{\rho_n}\kappa\rho_n, \eta_{\rho_n} - \kappa\rho_n, \mathbb{S}_{\rho_n}).$$

Working similarly to the above case, we get

$$(6.15) \quad 0 \geq - \left(\frac{\partial \phi(X_0)}{\partial y} - (1-\mu)S_0 \frac{\partial \phi(X_0)}{\partial B} \right).$$

Case 3. It holds that

$$\mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}) = \mathbb{E}\{\mathbb{V}_j^{\rho_n}(\iota_{\rho_n} + \rho_n, \mathbb{B}_{\rho_n} \exp(r\rho_n), \eta_{\rho_n}, \mathbb{S}_{\rho_n} \exp(\alpha\rho_n + \theta\sigma\sqrt{\rho_n}))\}.$$

Then (6.10) implies that

$$\begin{aligned}
 (6.16) \quad & \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}) \\
 & \geq \mathbb{E}\{\phi(\iota_{\rho_n} + \rho_n, \mathbb{B}_{\rho_n} \exp(r\rho_n), \eta_{\rho_n}, \mathbb{S}_{\rho_n} \exp(\alpha\rho_n + \theta\sigma\sqrt{\rho_n}))\} \\
 & + \mathbb{V}_j^{\rho_n}(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}) - \phi(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n}),
 \end{aligned}$$

and therefore

$$\begin{aligned}
 0 & \geq \liminf_n \left\{ \frac{\phi(\iota_{\rho_n} + \rho_n, \mathbb{B}_{\rho_n} \exp(r\rho_n), \eta_{\rho_n}, \mathbb{S}_{\rho_n} \exp(\alpha\rho_n + \theta\sigma\sqrt{\rho_n})) - \phi(\iota_{\rho_n}, \mathbb{B}_{\rho_n}, \eta_{\rho_n}, \mathbb{S}_{\rho_n})}{\rho_n} \right\} \\
 (6.17) \quad & \geq \liminf_{\rho \rightarrow 0} \left\{ \frac{\phi(t_0 + \rho, B_0 \exp(r\rho), y_0, S_0 \exp(\alpha\rho + \theta\sigma\sqrt{\rho})) - \phi(t_0, B_0, y_0, S_0)}{\rho} \right\} \\
 & = \frac{\partial \phi(X_0)}{\partial s} + rB_0 \frac{\partial \phi(X_0)}{\partial B} + \alpha S_0 \frac{\partial \phi(X_0)}{\partial S} + \frac{1}{2} \sigma^2 S_0^2 \frac{\partial^2 \phi(X_0)}{\partial S^2}.
 \end{aligned}$$

Combining (6.14), (6.15), and (6.17) yields (6.11), and the proof is complete. \square

In the discrete time framework, the exponential utility function, given by (4.21), has the same effect on the value function as before, and therefore (4.25) can be written as follows:

$$(6.18) \quad \mathbb{V}_j(\iota, \mathbb{B}, \eta, \mathbb{S}) = 1 - \exp\left(-\gamma \frac{\mathbb{B}}{\Delta(N, \iota)}\right) \mathbb{Q}_j(\iota, \eta, \mathbb{S}),$$

where $\Delta(\nu, \nu')$ is the discrete time discount factor, given by

$$(6.19) \quad \Delta(\nu, \nu') = \exp(-r(\nu - \nu')),$$

where ν and ν' take values in the same set with ι and $\nu \geq \nu'$. The discretisation scheme for the new functions $\mathbb{Q}_j(\iota, \eta, \mathbb{S}) := 1 - \mathbb{V}_j(\iota, 0, \eta, \mathbb{S})$ is derived from (6.4) and (6.18) to be

$$(6.20) \quad \begin{aligned} \mathbb{Q}_j(\iota, \eta, \mathbb{S}) = \min \{ & \mathbb{F}_b(\iota, \zeta, \mathbb{S}) \times \mathbb{Q}_j(\iota, \eta + \zeta, \mathbb{S}), \\ & \mathbb{F}_s(\iota, \zeta, \mathbb{S}) \times \mathbb{Q}_j(\iota, \eta - \zeta, \mathbb{S}), \mathbb{E}\{\mathbb{Q}_j(\iota + \zeta, \eta, \Theta \times \mathbb{S})\} \}, \end{aligned}$$

where

$$(6.21) \quad \mathbb{F}_b(\iota, \zeta, \mathbb{S}) = \exp\left(\gamma \frac{(1 + \lambda)\mathbb{S}\zeta}{\Delta(N, \iota)}\right)$$

and

$$(6.22) \quad \mathbb{F}_s(\iota, \zeta, \mathbb{S}) = \exp\left(-\gamma \frac{(1 - \mu)\mathbb{S}\zeta}{\Delta(N, \iota)}\right),$$

and the boundary conditions at $\iota = N$ are given by the discrete space versions of (4.28) and (4.29). As in the continuous time case, if the value functions are known in the no-transaction region, then they can be calculated in the buy and sell regions by using the discrete space versions of (4.15) and (4.17): Suppose that η_b^* is the value of η , at which it is optimal to buy ζ shares, whereas, at $\eta = \eta_b^* + \zeta$ it is optimal to perform no transaction at all; then the function $\mathbb{Q}_j(\iota, \eta, \mathbb{S})$ is determined by

$$(6.23) \quad \mathbb{Q}_j(\iota, \eta, \mathbb{S}) = \mathbb{F}_b(\iota, \eta_b^* - \eta, \mathbb{S}) \times \mathbb{Q}_j(\iota, \eta_b^*, \mathbb{S}) \quad \forall \eta < \eta_b^*,$$

and a similar equation can be derived for $\eta > \eta_s^*$, the value of η at which it is optimal to sell ζ shares by using $\mathbb{F}_s(\iota, \eta - \eta_s^*, \mathbb{S})$. Finally, the price of the European contingent claim is given by

$$(6.24) \quad \mathbb{P}_w(\iota, \mathbb{S}) = \frac{\Delta(N, t)}{\gamma} \ln\left(\frac{\mathbb{Q}_w(\iota, 0, \mathbb{S})}{\mathbb{Q}_1(\iota, 0, \mathbb{S})}\right),$$

which is the discrete time version of (4.26).

7. Numerical results. The algorithm developed in the previous section was implemented, and the writing price of a European call option was calculated for a writer with exponential utility function given by (4.21) and for constant model coefficients. For comparison, the Black-Scholes value was also calculated from

$$(7.1) \quad p_{bs}(s, S) = SN(\chi) - Ee^{-r(T-s)}N(\chi - \sigma\sqrt{T-s}),$$

where

$$(7.2) \quad \chi = \frac{\ln(S/Ee^{-r(T-s)})}{\sigma\sqrt{T-s}} + \frac{1}{2}\sigma\sqrt{T-s},$$

and $N(\cdot)$ denotes the normal distribution function with mean zero and variance 1. The number of stock units held in the hedging portfolio is simply $N(\chi)$. The boundary conditions for the value functions were set according to the analysis presented in § 1, and several values of proportional transaction costs were tried. The effects on p_w of the model parameters, the time to expiration, and the stock price at the time the option is written were in line with expectations and are outlined below. The parameters λ and μ were set equal to each other and are denoted by T.C. in the figures.

For both value functions, the boundaries ∂B and ∂S were found to lie below and above the optimal trading strategy without transaction costs, and the no-transaction region was observed to widen as the expiration date approached. (Note that we have not proved that the optimal transaction policies consist of reflection off these boundaries, although, of course, we believe this to be the case.) This shows that the investor is reluctant to transact toward the end of the trading interval, as he thinks that the stock price may not vary too much until the final time. (The cost of transactions is likely to reduce the utility of the final wealth more than the cost of providing one share at the final time.) Also, the boundary ∂B , for the value function $V_w(s, B, y, S)$, was virtually equal to the Black-Scholes trading strategy $N(\chi)$, indicating that the writer considers the cost of the obligation to provide one share of stock at the final time (if he does not already own it) as the most significant factor, affecting the trading strategy for this value function.

The most important result appears in Fig. 1, where the price difference $p_w - p_{bs}$ is plotted against time over a three-year period, with the parameter values for a one-year period shown in its title (p_{bs} is the Black-Scholes price, given by (7.1)). The expiration date is at the end of the third year, where both prices vanish, as the option is worthless. The price difference a long time before the expiration of the claim is equal to λS , which is the amount required to buy one share of the stock. The reason for this is revealed by observing $N(\chi)$ over the three time periods, plotted in Fig. 2 for the same parameter values. Although $S < E$, $N(\chi)$ increases as the time to expiration increases and dictates that the investor must own almost one share of the stock in a market with no transaction costs if the time to expiration is long enough. By that time, the extra price that the writer charges is the amount required to buy one share of the stock,

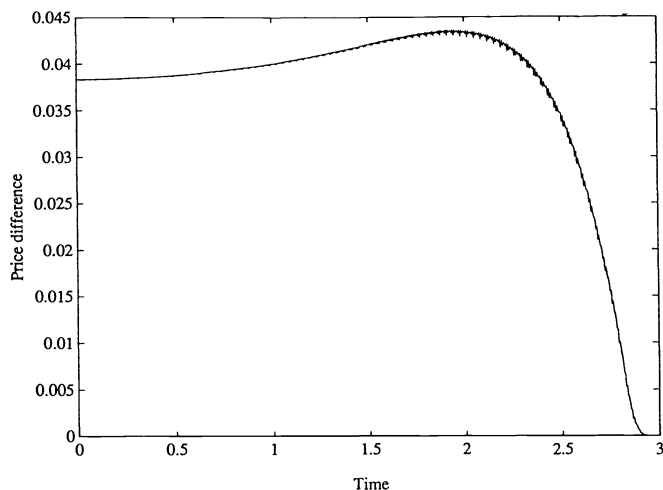
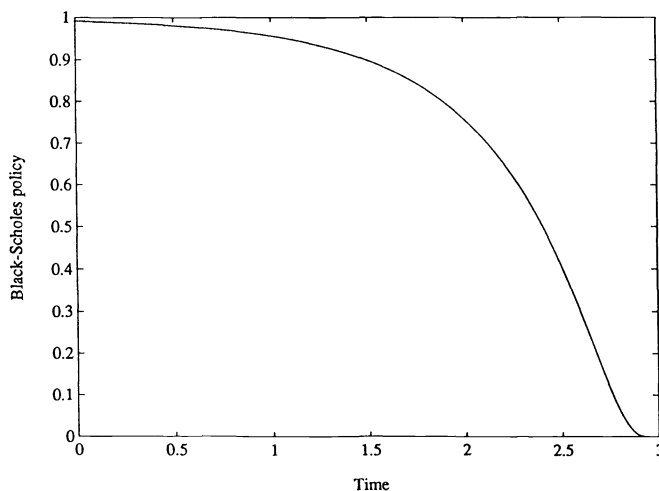
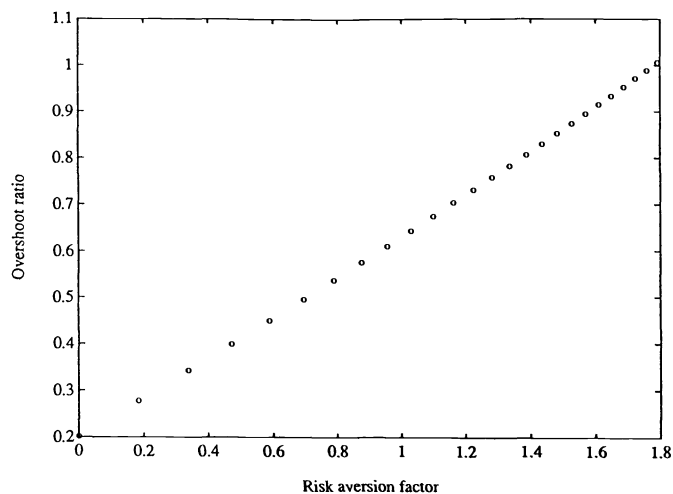


FIG 1. T.C. = 0.2 percent; $S = 19$; $E = 20$; $\gamma = 1.0$; $\sigma = 0.05$; $r = 8.5$ percent; $\alpha = 10$ percent.

FIG 2. $S = 19$; $E = 20$; $\sigma = 0.05$; $r = 8.5$ percent.

which is the “hedging” strategy of the option writer, who performs few transactions long before expiration because of the wide range of possible paths of the stock price until the final time. Also, as the final time approaches, the price difference shows a “hump,” which represents the period of active trading by the option writer. Finally, if $S > E$, the price difference is λS at the final time.

The variation of the peak of the price difference with the model parameters was investigated, and the following results were obtained. The “overshoot” ratio $((p_w - p_{b_s}) - \lambda S) / \lambda S$ was calculated, and it was observed that it is (i) a linear increasing function of the logarithm of the index of risk aversion γ (ii) a linear increasing function of the volatility σ , (iii) a decreasing function of the stock price S , (iv) a convex decreasing function of the proportional transaction charge $\lambda = \mu$, (v) a convex function of the interest rate r , and (vi) a linear decreasing function of the stock’s mean growth rate α . These results are illustrated in Figs. 3–8, where the relevant parameter values

FIG. 3. T.C. = 0.2 percent; $S = 9$; $E = 10$; $\sigma = 0.075$; $r = 7$ percent; $\alpha = 10$ percent.

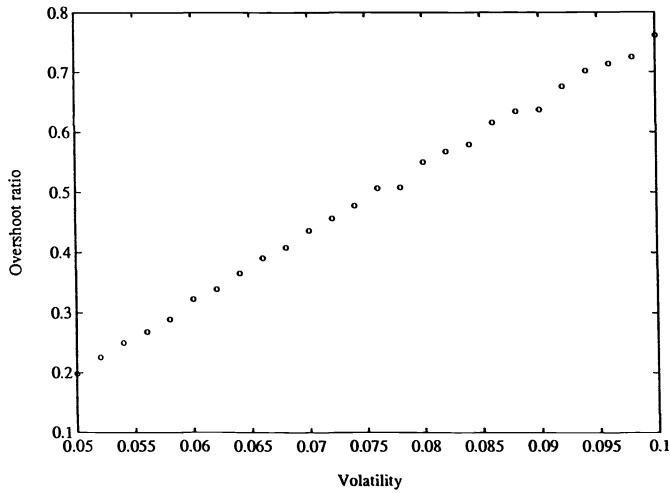


FIG. 4. T.C. = 0.2 percent; $S = 29$; $E = 30$; $\gamma = 1.0$; $r = 7$ percent; $\alpha = 10$ percent.

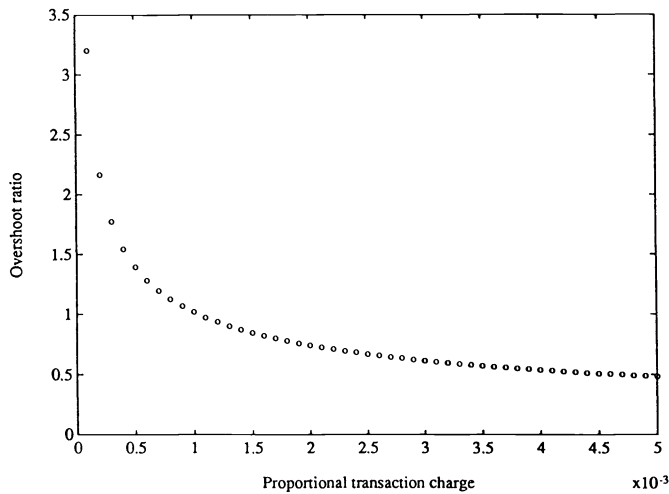


FIG. 5. $S = 29$; $E = 30$; $\gamma = 2.0$; $\sigma = 0.075$; $r = 7$ percent; $\alpha = 10$ percent.

are shown in their titles. The results are interpreted as follows. As the writer becomes more risk averse, the boundaries ∂B and ∂S come closer to the optimal trading strategy without transaction costs, thus mandating more transactions and increasing the option price. The linearity with $\ln(\gamma)$ is probably due to the form of the utility function. As the volatility of the underlying risky security increases, the uncertainty facing the option writer is greater, and the option price increases, as is in the case without transaction costs. As the stock price increases over the exercise price, the price difference is λS ; at expiration, the Black-Scholes strategy $N(\chi)$ dictates holding one share of the stock until expiration, and the “hump” is small. As the transaction costs increase, λS increases; but the above ratio decreases as the writer tries to perform less transactions (the boundaries ∂B and ∂S move away from the optimal trading strategy without transaction costs). Finally, the mean growth rate α and the interest rate r have little effect on the above ratio as compared to other parameters.

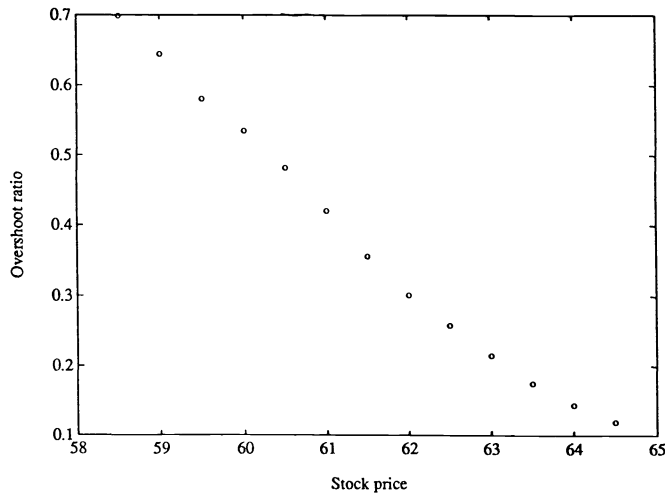


FIG. 6. T.C. = 0.2 percent; $E = 60$; $\gamma = 1.0$; $\sigma = 0.075$; $r = 7$ percent; $\alpha = 10$ percent.

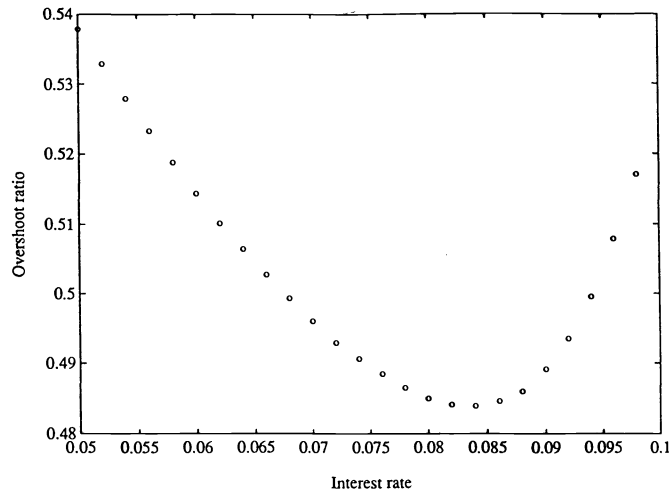


FIG. 7. T.C. = 0.2 percent; $S = 29$; $E = 30$; $\gamma = 1.0$; $\sigma = 0.075$; $\alpha = 10$ percent.

8. Concluding remarks. There are several directions in which our approach needs further investigation.

1. *Nonexponential utilities.* There is no issue of principle here, but only of a further increased computational load, since the reduction from four dimensions to three is no longer available. Since the risk averse writer's strategy is basically a hedging strategy, we believe that the form of the utility function is unimportant and that only its curvature at the origin plays any real role. If true, this would provide a justification for using the computationally simpler exponential function.

2. *Diversified portfolios.* As pointed out earlier, in our framework the writer may well wish to include other stocks (not just the one on which the option is written) in the hedging portfolio. Again, this simply increases the dimensionality of the problem. By allowing investment in other stocks we are enlarging the class of possible hedging strategies, and hence the option writing price will be reduced; we do not know, however, by how much.

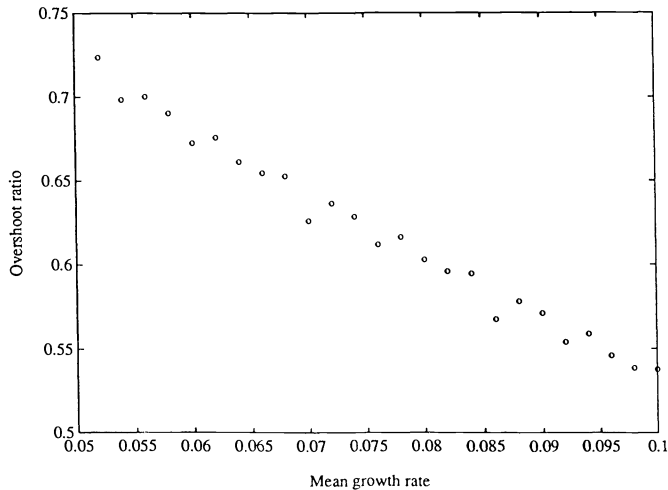


FIG. 8. T.C. = 0.2 percent; $S = 29$; $E = 30$; $\gamma = 1.0$; $\sigma = 0.075$; $r = 5$ percent.

3. *American options.* Clearly, a similar approach could be taken to the pricing of American options. There is, however, a conceptual problem in that the buyer, not the writer, controls the exercise strategy, and the pricing problem must involve the solution of one more utility maximisation problem over all the exercising strategies available to the buyer. In particular, there seems no reason why the buyer should use the “frictionless exercise strategy” as described in § 6 of Karatzas [13]. The precise definition of the problem is currently under investigation.

4. *Equilibrium.* Under what circumstances will a writer and a buyer agree on a “deal,” i.e., a common price for an option contract in the framework we have described? This is a very important question; one that we do not claim to understand fully. It is possible to define a buying price in a way that mirrors our definition of the writing price, and this is what Hodges and Neuberger [11] do. In the notation of § 2, if the buyer forms a hedging portfolio whose composition at the exercise time T is (B, \underline{y}) then its cash value after exercise of the option is $B - E + c(\underline{y} + \underline{e}_1, \underline{S})$. Analogous to the definition (2.1) for V_w , we can therefore define

$$(8.1) \quad V_b(B) = \sup_{\pi \in \mathcal{T}(B)} \mathbb{E} \{ \mathcal{U}(\mathbf{B}^\pi(T) + I_{(S_1(T) \leq E)} c(\underline{y}^\pi(T), \underline{S}(T)) + I_{(S_1(T) > E)} [c(\underline{y}^\pi(T) + \underline{e}_1, \underline{S}(T)) - E] \}$$

and

$$(8.2) \quad B_b = \inf \{ B : V_b(B) \geq 0 \},$$

and the buying price p_b as

$$(8.3) \quad p_b = B_b - B_1,$$

where B_1 is given by (2.4). However, we do not believe this definition to be an appropriate one. The most obvious objection is that it is very hard to see how writer and buyer could ever agree on a price. Certainly $p_w \neq p_b$ if all parameters are the same for all calculations. One may hypothesise that writer and buyer agree on market parameters, but have different utility functions. This fails, however, because it is always the case that $p_w > p_{bs}$ and $p_b < p_{bs}$, where p_{bs} is the (preference-independent) Black-Scholes price. At a more fundamental level the above buying price seems inappropriate

because it fails to respect the essential asymmetry in an option contract, namely that buying an option is a form of insurance, whereas writing one is a gamble. This distinction disappears in the Black-Scholes world, as there is no essential element of risk on either side, but no general theory of option pricing can be satisfactory if this distinction is not taken into account. This is an interesting area for further research.

Acknowledgments. The starting point for this work was a preliminary version of the paper [11], presented by Stewart Hodges and Anthony Neuberger at a workshop organised by the Financial Options Research Centre of Warwick University in 1989. Readers of [11] will appreciate that several of the key ideas were introduced in that paper. Collaborative work on the present paper was initiated while the first and third authors were visiting the Graduate School of Business, Stanford University, December 1990. We would like to thank Darrell Duffie, Avi Mandelbaum and the Finance and Decision Sciences groups at the Graduate School of Business at Stanford University for this opportunity as well as for stimulating comments on the work in progress. We also appreciate some incisive comments by Lucien Foldes at the London School of Economics and Political Science and discussions with Helyette Geman, which clarified our thinking.

REFERENCES

- [1] G. BARLES AND P. E. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal., 4 (1991), pp. 271-283.
- [2] B. BENSARD, J. LESNE, H. PAGES, AND J. SCHEINKMAN, *Derivative asset pricing with transaction costs*, working paper, Bank of France, Centre de Recherche (1991).
- [3] F. BLACK AND M. SCHOLES, *The pricing of options and corporate liabilities*, J. Political Economy, 81 (1973), pp. 637-659.
- [4] CAPUZZO-DOLCETTA, AND P.-L. LIONS, *Hamilton-Jacobi equations and state constraints problems*, IMA preprint ser. 342, University of Minnesota, MN, 1987.
- [5] J. C. COX AND M. RUBINSTEIN, *Options Markets*, Prentice-Hall, Englewood Cliffs, NJ, 1985.
- [6] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.
- [7] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Department of Mathematics, University of California, Santa Barbara, 1990, preprint.
- [8] M. H. A. DAVIS AND A. R. NORMAN, *Portfolio selection with transaction costs*, Math. Oper. Res., 15 (1990), pp. 676-713.
- [9] M. H. A. DAVIS AND V. G. PANAS, *European option pricing with transaction costs*, Proc. 30th I.E.E.E. Conference on Decision and Control (1991), pp. 1299-1304.
- [10] C. EDIRISINGHE, V. NAIK, AND R. UPPAL, *Optimal replication of option with transaction costs*, University of British Columbia, working paper, 1991.
- [11] S. D. HODGES AND A. NEUBERGER, *Optimal replication of contingent claims under transaction costs*, Rev. Futures Markets, 8 (1989), pp. 222-239.
- [12] H. ISHII AND P.-L. LIONS, *Viscosity solutions of fully non-linear second-order elliptic partial differential equations*, J. Differential equations, 83, 1 (1990), pp. 26-78.
- [13] I. KARATZAS, *Optimisation problems in the theory of continuous trading*, SIAM J. Control Optim., 27 (1989), pp. 1221-1259.
- [14] M. KATSOUKAKIS, *State-constraint problems for second order fully nonlinear degenerate partial differential equations*, Ph.D. thesis, Brown University, Providence, RI, 1991.
- [15] L. H. MARTINS AND H. J. KUSHNER, *Numerical methods for stochastic singularly controlled problems*, SIAM J. Control Optim., 29 (1991), pp. 1443-1475.
- [16] H. E. LELAND, *Option pricing and replication with transactions costs*, J. Finance, 40 (1985), pp. 1283-1301.
- [17] P.-L. LIONS, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations*, parts 1 & 2, Comm. Partial Differential Equations, 8 (1983), pp. 1101-1174, 1229-1276.
- [18] H. M. SONER, *Optimal control with state-space constraints*, SIAM J. Control Optim., 24 (1986), pp. 552-561.
- [19] H. ZHU, *Characterisation of variational inequalities in singular control*, Ph.D. thesis, Brown University, Providence, RI, 1991.

ESTIMATION OF THE QUADRATIC VARIATION OF NEARLY OBSERVED SEMIMARTINGALES WITH APPLICATION TO FILTERING*

JEAN PICARD†

This paper is dedicated to Wendell Fleming on the occasion of his 65th birthday.

Abstract. Consider a filtering problem in which the available information is a noisy observation of a continuous semimartingale H_t . In the case of a high signal-to-noise ratio, it is proved that H_t and its quadratic variation can be jointly estimated by means of a finite-dimensional filter; moreover, for this result, the observation noise and H_t are not required to be independent. This problem can be viewed as a linear filtering problem with randomly time-varying parameters, and our filter is auto-adaptive with respect to changes of the parameters. These results are then applied to the nonlinear filtering of Markov diffusion processes when the observation function is not injective but satisfies a weaker detectability assumption. It appears that filtering such a system involves two timescales. The study is based on time discretization; the main tools are an averaging principle and an application of the asymptotic ordinary differential equation method for the study of stochastic algorithms.

Key words. nonlinear filtering, systems with small noise, linear filtering with unknown parameters, adaptive filtering, averaging principle, stochastic algorithms

AMS(MOS) subject classifications. 93E11, 60G35, 62F35

Introduction. The problem of nonlinear filtering consists in estimating some stochastic process that is not directly observed. Recently, many works were devoted to the case where the process is nearly observed; this is indeed an asymptotic framework in which several computable filters can be mathematically studied. More precisely, suppose that the observation is the d -dimensional process Y_t , which is given by

$$(0.1) \quad Y_t = \int_0^t H_s ds + \varepsilon B_t,$$

where H_t is an adapted continuous process, B_t is a standard Wiener process independent of H_t and ε is a nonnegative parameter. If $\varepsilon = 0$, then H_t is adapted to the filtration \mathcal{Y}_t generated by Y_t ; we will say that H_t is observable. Similarly, any process Z_t that is adapted to the filtration of H is observable. If ε is positive and small, we may expect that such a process is asymptotically observable; we can indeed prove with the method of [17] that, if Z_t is integrable, then its conditional mean, given \mathcal{Y}_t , denoted by \hat{Z}_t , converges to Z_t in L^1 as ε tends to zero. However, the exact computation of \hat{Z}_t generally requires us to solve an infinite-dimensional equation, so we look for a computable approximation. This means that we must find an observable process \bar{Z}_t , which is obtained by solving a finite-dimensional equation and such that $\|Z_t - \bar{Z}_t\|_1$ tends to zero as $\varepsilon \rightarrow 0$. Another problem consists in estimating $\hat{Z}_t - \bar{Z}_t$; it has at most the order of $Z_t - \bar{Z}_t$, but can be much smaller if the filter is efficient.

An important case is the filtering of diffusion processes; we suppose that $H_t = h(X_t)$, where X_t is a n -dimensional diffusion process given by

$$(0.2) \quad dX_t = f(X_t)dt + g(X_t)dW_t.$$

* Received by the editors November 22, 1991; accepted for publication (in revised form) June 10, 1992.

† Laboratoire de Mathématiques Appliquées (CNRS-URA 1501), Université Blaise Pascal (Clermont-Ferrand II), 63177 Aubière Cedex, France.

Assuming that the smooth coefficients (f, g, h) are such that X_t is adapted to the natural filtration of $h(X_t)$ (this is a detectability assumption), we want to design an efficient suboptimal filter. The easiest case is the case where h is injective; see [5], [12], [1] for the scalar case and [13], [14] for the multidimensional case. Under some additional conditions, it can be proved that the extended Kalman filter provides an observable process \bar{X}_t such that $X_t - \bar{X}_t$ and $\hat{X}_t - \bar{X}_t$ are, respectively, of order $\sqrt{\varepsilon}$ and ε ; this filter requires us to solve an equation of dimension $n(n+3)/2$, but we can also find a simplified filter of dimension n satisfying the same estimates. Moreover, a deviation in the filter at time t (caused, for instance, by an aberrant observation) is forgotten after time t with an exponential rate of order $1/\varepsilon$; we say that the memory length is of order ε (some formal definitions of the memory length are proposed in [10]). More recently, conditions that are more general than the injectivity of h have been considered; in these cases, we can again find approximate filters, but $X_t - \bar{X}_t$ is generally larger than $\sqrt{\varepsilon}$. A first example is the case where $h'g = 0$ and $x \mapsto (h(x), Lh(x))$ injective, where h' is the Jacobian matrix of h and L is the generator of x ; this means that H_t is absolutely continuous and that X_t is a function of (H_t, \bar{H}_t) . A class of systems entering into this framework is studied in [18] and [11, §3], and we find a filter \bar{X}_t such that $X_t - \bar{X}_t$ is of order $\varepsilon^{1/4}$, $h(X_t) - h(\bar{X}_t)$ is of order $\varepsilon^{3/4}$, and the memory length is of order $\sqrt{\varepsilon}$. Thus the components of X_t are observed with different precisions but in the same timescale.

A second example is the case where $x \mapsto (h(x), h'gg^*h'^*(x))$ injective; if we let $\langle H, H \rangle_t$ be the quadratic covariation matrix of the process H_t , this assumption means that X_t is a function of H_t and $d\langle H, H \rangle_t/dt$, so that X_t is again adapted to the filtration of H_t . Note that, contrary to previous examples, this case is purely nonlinear, in the sense that a linear system cannot satisfy this injectivity condition except when h itself is injective. First, suppose that, for any h_0 , the set $\{h(x) = h_0\}$ is finite. Since we have a good observation of $h(X_t)$, then, to find an observable approximation of X_t , we must only choose between a finite number of possible approximations. This can be achieved by one of the test procedures described in [3]. Our results can be applied to the design of such a test, but we instead concentrate on the case where the sets $\{h(x) = h_0\}$ are connected manifolds. Then the filtering problem can be decomposed into, first, an estimation \bar{H}_t of $h(X_t)$ and, second, an estimation \bar{X}_t of X_t on the manifold $\{h(x) = \bar{H}_t\}$; a particular example has already been studied in [15]. Here, we find a family of observable processes \bar{X}_t , which can be computed by solving finite-dimensional equations and which approximate X_t . Their infinitesimal increment $d\bar{X}_t$ is obtained by adding a component due to the variations of $h(X_t)$ and a component that is tangent to $\{h(x) = h(\bar{X}_t)\}$ and that is due to the variations of the other components of X_t . Roughly speaking, these filters look like Kalman filters, but their gain processes are obtained from stochastic differential equations rather than Riccati equations. It is proved that $X_t - \bar{X}_t$ and $h(X_t) - h(\bar{X}_t)$ are, respectively, of order $\varepsilon^{1/4}$ and $\sqrt{\varepsilon}$, that $h(\bar{X}_t) - \bar{H}_t$ is of order $\varepsilon^{3/4}$, and that the memory length is of order ε for the estimation of $h(X_t)$, but only of order $\sqrt{\varepsilon}$ for the other components. Thus, the filter has two timescales; for detectable almost linear systems, it was proved in [6] that multitimescale filters are involved when the noises have different levels on the various components of X_t and Y_t . Here, multitimescale properties are a consequence of the geometric structure of the system.

Until now, we have explained the nature of our results for the filtering of diffusion processes; however, we will first consider this problem in a non-Markovian setting. This situation is much more general with regard to the assumptions, but we consider it easier to understand because we have not taken into account the geometric structure of the diffusion process. We also allow the signal and observation noises to be correlated.

We suppose that (H_t, Y_t) is an $(\mathbb{R}^d \times \mathbb{R}^d)$ -valued Itô process of the form

$$(0.3) \quad H_t = H_0 + \int_0^t F_s ds + \int_0^t G_s dW_s,$$

$$(0.4) \quad Y_t = \int_0^t H_s ds + \varepsilon \int_0^t J_s dW_s$$

for a standard m -dimensional Wiener process W_t and adapted processes F_t , G_t , and J_t . The problem of estimating H_t can be viewed as a linear filtering problem with time-varying unknown parameters (F_t, G_t, J_t) . The process H_t can be estimated with an error of order $\sqrt{\varepsilon}$ and a memory length of order ε , and we want to know which part of the parameters can be estimated. The Markovian case described above is then obtained by putting some nonlinear constraints on H_t and the parameters. First, it is clear that, by computing the quadratic variation of Y_t , $J_t J_t^*$ can be estimated with an arbitrary precision. On the other hand, F_t cannot generally be estimated with a small error. Take, for instance, $J_t = 0$, $G_t = I$, and $F_t = F$; then we observe a Wiener process with drift F , but it is well known that the drift can be identified only after a large amount of time. In this paper, we prove that $G_t G_t^*$ can be estimated with an error of order $\varepsilon^{1/4}$ and a memory length of order $\sqrt{\varepsilon}$; concerning the correlation between the noises, it appears that the skew-symmetric part of $G_t J_t^*$ can be estimated in the same conditions, but its symmetric part is generally not asymptotically observable. These results are proved by discretizing the time with the step $\varepsilon^{3/4}$, which is intermediate between the two timescales; thus, with respect to the phenomena occurring in the timescale ε , we can prove an averaging principle, and, with respect to the phenomena occurring in the timescale $\sqrt{\varepsilon}$, the discretized problem looks like the long-time behaviour of a stochastic algorithm with a small slowly varying gain. There exists a wide literature about stochastic algorithms, see, for instance [2], but our exposition will be self-contained. The main idea is that the behaviour of these algorithms can be approximately described by an ordinary differential equation. This tool was already applied in [9] and [16] to the estimation of fixed parameters in a linear system; here, the parameters are not constant but vary slowly with respect to the memory length of the filter. Observe, in particular, that the limiting differential equation involves random coefficients.

The approximate knowledge of the parameters enables us to improve the estimation of \hat{H}_t , so the next step consists of designing an adaptive filter that jointly estimates \hat{H}_t and the parameters. This can be done under stronger conditions, and it appears that \hat{H}_t can be computed with an error of order $\varepsilon^{3/4}$, even if the parameters cannot be identified; we say that the filter is self-tuning. The next step, which is not dealt with here, would consist in comparing the efficiency of the different filters concerning the parameter estimation; for this purpose, we must observe that the filtering problem actually contains two periods: First, we must estimate quickly and roughly the system for small times; if we have no knowledge on the initial conditions, this requires uniformly stable filters. Then, starting with these rough estimates, we must track the system as it evolves; as long as we neglect large deviations phenomena, this requires only locally stable filters. If we want to optimize the filter with respect to its local tracking properties, we obtain a filter that is locally (but not necessarily uniformly) stable. Thus we sometimes must use two different filters for the two periods. In this paper, we describe a filter that can be proved uniformly stable, and we introduce a wider class of locally stable filters.

These algorithms are then applied to the local tracking of Markovian diffusions observed with a small correlated noise. We suppose that the n -dimensional signal X_t

and the d -dimensional observation $Y_t (d < n)$ satisfy

$$(0.5) \quad dX_t = f(X_t) dt + g(X_t) dW_t,$$

$$(0.6) \quad dY_t = h(X_t) dt + \varepsilon j(X_t) dW_t.$$

We also suppose that jj^* and $h'g(I-j^*j)g^*h'^*$ are elliptic; this last condition means that there is some nondegenerate noise in $h(X_t)$ that is independent of the observation noise. If it is not satisfied, the filtering error may be too small for our method to work. Under these assumptions, the linear system

$$(0.7) \quad dH_t = h'g(x) dW_t, \quad dY_t = H_t dt + j(x) dW_t,$$

obtained by freezing the coefficients at x and normalizing the noise, admits a stationary Kalman filter, and we denote by $p(x)$ its gain. Then our local detectability condition is that $x \mapsto (h(x), p(x))$ is locally injective; under this condition and if X_0 is known, we can find a good filter \bar{X}_t that tracks X_t . Under the additional assumption $jj^* = I$, which implies that the filtering problem is not singular, we prove that $h(\bar{X}_t)$ is a good approximation of the conditional expectation of $h(X_t)$ given \mathcal{Y}_t . Observe that if, moreover, the noises are independent (this means that $fg^* = 0$), we have $p = (h'gg^*h'^*)^{1/2}$.

Let us now briefly describe the contents of this work. In § 1 we first set our notation. In § 1.1 we describe filters that do not necessarily give the best results but that work in a quite general framework. In § 1.2, under some additional assumptions, we describe some adaptive filters, and in § 1.3 we consider the Markovian case. These results are respectively proved in §§ 2–4.

1. The results. Throughout this work, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration \mathcal{F}_t , and we let W_t be a standard m -dimensional \mathcal{F}_t -Wiener process. We suppose that \mathcal{F}_0 is trivial; this condition is not a restriction (if it is not satisfied, we can apply our results by working conditionally on \mathcal{F}_0) and is only used for notational convenience. It means that the initial condition of our system is unknown but deterministic. We also have a filtration $\mathcal{Y}_t \subset \mathcal{F}_t$ generated by an observation process Y_t , and \mathcal{Y}_t -adapted processes are said to be observable.

Let us set some definitions. If Z_t is a family (indexed by ε) of stochastic processes and if z_t is a family of deterministic positive functions, we say that Z_t is (at most) of order z_t , and we write $Z_t = O(z_t)$ if, for any $1 \leq k < \infty$,

$$(1.1) \quad \limsup_{\varepsilon \rightarrow 0} \sup_t \frac{\|Z_t\|_k}{z_t} < \infty$$

for t in some bounded or unbounded time interval (which will be made precise in the statements). If U is a family of events, we say that Z_t is of order z_t on U if $Z_t 1_U$ is of order z_t . The statement $Z_t - Z_s = O(\varepsilon^\beta)$ for $|t-s| \leq \varepsilon^\alpha$ means that

$$(1.2) \quad \limsup_{\varepsilon \rightarrow 0} \sup \{ \varepsilon^{-\beta} \|Z_t - Z_s\|_k ; |t-s| \leq \varepsilon^\alpha \} < \infty$$

for any k . We sometimes use the weaker notion of domination in probability; we say that Z_t is of order z_t in probability if

$$(1.3) \quad \lim_{\lambda \rightarrow \infty} \limsup_{\varepsilon \rightarrow 0} \sup_t \mathbb{P}[|Z_t| > \lambda z_t] = 0.$$

In particular, if Z_t is $O(\varepsilon^k)$ on an event of probability tending to 1, then Z_t is of order ε^k in probability. The notation $O(\varepsilon^\infty)$ means of order ε^k for any k . The set M_d is the space of $d \times d$ matrices; it is a Euclidean space with the product $\text{trace}(AB^*)$. The

subset of matrices, the eigenvalues of which have positive real part, will be denoted by M_d^+ . For A in M_d^+ , we put

$$(1.4) \quad A^\dagger = 2 \int_0^\infty e^{-A^*s} e^{-As} ds.$$

Observe that, if A is symmetric, A^\dagger is its inverse. A family of processes Z_t is said to be of class \mathcal{J} if Z_t are Itô processes such that $Z_0 = O(1)$ and

$$(1.5) \quad dZ_t = O(1)dt + O(1)dW_t.$$

For $\alpha \in \mathbb{R}$, we say that a family of M_d -valued processes A_t is exponentially stable in the timescale ε^α if the solution of

$$(1.6) \quad Z_t = I + \varepsilon^{-\alpha} \int_0^t A_s Z_s ds$$

satisfies

$$(1.7) \quad |Z_t Z_s^{-1}| \leq C \exp \{-c\varepsilon^{-\alpha}(t-s)\}$$

for some positive c and C . Examples of such processes are $A_t = -P_t$, where P_t is symmetric and uniformly elliptic, or $A_t = -P$, where P is in a compact subset of M_d^+ ; in these two examples, A_t is exponentially stable in any timescale. Other conditions ensuring the exponential stability will be given in Lemma 1.

1.1. Estimation of the signal and its quadratic variation. Consider the process (H_t, Y_t) defined by (0.3), (0.4) for some adapted processes F_t, G_t, J_t , which may depend on ε . We want to find observable approximations of the processes H_t, G_t , and J_t , assuming the knowledge of $J_t J_t^*$. The filter for H_t is given by

$$(1.8) \quad d\bar{H}_t = \bar{F}_t dt + \frac{\bar{P}_t}{\varepsilon} (dY_t - \bar{H}_t dt)$$

for some drift \bar{F}_t , some gain process \bar{P}_t , and some initial condition \bar{H}_0 .

THEOREM 1. Assume that H_t is of class \mathcal{J} , that $\bar{F}_t = O(1)$, that $J_t = O(1)$, and that $-\bar{P}_t$ is bounded and exponentially stable in the timescale ε . Then

$$(1.9) \quad H_t - \bar{H}_t = O(\sqrt{\varepsilon} + |H_0 - \bar{H}_0| e^{-ct/\varepsilon})$$

for some $c > 0$.

As is often the case for systems with small observation noise, the value of \bar{F}_t is not important for the efficiency of the filter. However, the dependence of \bar{H}_t on the process \bar{P}_t is crucial; the estimation of the parameters G_t and J_t is based on differentiation with respect to \bar{P}_t . For $E \in M_d$, let $\Gamma_t^E = \nabla_E \bar{H}_t$ be the derivative of \bar{H}_t with respect to a perturbation of \bar{P}_t in the direction E . We have

$$(1.10) \quad d\Gamma_t^E = -\frac{\bar{P}_t}{\varepsilon} \Gamma_t^E dt + \frac{E}{\varepsilon} (dY_t - \bar{H}_t dt)$$

with $\Gamma_0^E = 0$. The map $E \mapsto \Gamma_t^E$ is linear from M_d into \mathbb{R}^d , and let us denote it by $\Gamma_t = \nabla \bar{H}_t$. In particular, its adjoint Γ_t^* is linear from \mathbb{R}^d into M_d , and we can consider the M_d -valued observable process $\int \Gamma_t^* (dY_t - \bar{H}_t dt)$. Note that

$$(1.11) \quad \int_s^t \Gamma_u^* (dY_u - \bar{H}_u du) \simeq \int_s^t \Gamma_u^* (H_u - \bar{H}_u) du = -\frac{1}{2} \nabla \int_s^t |H_u - \bar{H}_u|^2 du.$$

We now describe the asymptotic value of this process as $\varepsilon \rightarrow 0$. To this end, we need the following elementary results about linear systems.

For G, J in M_d and $\bar{P} \in M_d^+$, consider the linear system

$$(1.12) \quad dh_t = GdW_t, \quad dy_t = h_t dt + JdW_t, \quad d\bar{h}_t = \bar{P}(dy_t - \bar{h}_t dt),$$

and let $\Lambda(G, J, \bar{P})$ be the limit of $\mathbb{E}|h_t - \bar{h}_t|^2$ as $t \rightarrow \infty$; the function $\Lambda(G, J, \bar{P})$ is the trace of the matrix-valued function $\bar{Q} = \bar{Q}(G, J, \bar{P})$ defined by

$$(1.13) \quad \bar{P}\bar{Q} + \bar{Q}\bar{P}^* = (G - \bar{P}J)(G - \bar{P}J)^*.$$

By differentiating with respect to \bar{P} in the direction E , we obtain

$$(1.14) \quad \bar{P}\nabla_E \bar{Q} + \nabla_E \bar{Q}\bar{P}^* + E\bar{Q} + \bar{Q}E^* + EJ(G - \bar{P}J)^* + (G - \bar{P}J)J^*E^* = 0,$$

so

$$(1.15) \quad \nabla_E \bar{Q} = -\int_0^\infty e^{-\bar{P}s}(E\bar{Q} + \bar{Q}E^* + EJ(G - \bar{P}J)^* + (G - \bar{P}J)J^*E^*)e^{-\bar{P}^*s} ds.$$

By taking the trace, we obtain

$$(1.16) \quad \nabla_E \Lambda = -\text{trace}(\bar{P}^\dagger RE^*),$$

where \bar{P}^\dagger and $R = R(G, J, \bar{P})$ are defined by (1.4) and

$$(1.17) \quad R = \bar{Q} + GJ^* - \bar{P}JJ^*.$$

Thus,

$$(1.18) \quad \nabla \Lambda(G, J, \bar{P}) = -\bar{P}^\dagger R(G, J, \bar{P}).$$

Let us return to system (0.3), (0.4) with time-varying coefficients and to the filter (1.8). By means of an averaging principle and from the previous calculation, it will be proved in Lemma 2 that, for $t - s \gg \varepsilon$, under some conditions including $\bar{P}_t \in M_d^+$, we have

$$(1.19) \quad \int_s^t |H_u - \bar{H}_u|^2 du \simeq \varepsilon \int_s^t \Lambda(G_u, J_u, \bar{P}_u) du,$$

$$(1.20) \quad \nabla \int_s^t |H_u - \bar{H}_u|^2 du \simeq -\varepsilon \int_s^t \bar{P}_u^\dagger R_u du$$

with $R_t = R(G_t, J_t, \bar{P}_t)$, and therefore from (1.11)

$$(1.21) \quad \int_s^t \Gamma_u^*(dY_u - \bar{H}_u du) \simeq \frac{\varepsilon}{2} \int_s^t \bar{P}_u^\dagger R_u du.$$

Thus, we have an observable approximation of $\bar{P}_t^\dagger R_t$; since \bar{P}_t^\dagger is known, we can deduce approximations of R_t . If $K > 0$ and if \bar{R}_t is solution of

$$(1.22) \quad d\bar{R}_t = \frac{K}{\varepsilon^{3/2}} \left(\Gamma_t^*(dY_t - \bar{H}_t dt) - \frac{\varepsilon}{2} \bar{P}_t^\dagger \bar{R}_t dt \right),$$

it appears that $R_t - \bar{R}_t$ is of order $\varepsilon^{1/4}$ for $t \gg \sqrt{\varepsilon}$.

THEOREM 2. Assume the conditions of Theorem 1 and suppose, moreover, that $H_0 - \bar{H}_0$ is of order $\sqrt{\varepsilon}$, that G_t and J_t are of class \mathcal{F} , that \bar{P}_t takes its values in a compact subset of M_d^+ , and that

$$(1.23) \quad \bar{P}_t - \bar{P}_s = O(\varepsilon^{3/8}), \quad \mathbb{E}[\bar{P}_t - \bar{P}_s | \mathcal{F}_s] = O(\sqrt{\varepsilon})$$

for $|t-s| \leq \varepsilon^{3/4}$. For $K > 0$, let \bar{R}_t satisfy (1.22) with $K > 0$ and $\bar{R}_0 = O(1)$. Then

$$(1.24) \quad R_t - \bar{R}_t = O(\varepsilon^{1/4} + |R_0 - \bar{R}_0|e^{-ct/\sqrt{\varepsilon}})$$

uniformly for $t \geq 0$.

On the other hand, we check from (1.13) and (1.17) that

$$(1.25) \quad G_t G_t^* = \bar{P}_t R_t^* + R_t \bar{P}_t^* + \bar{P}_t J_t J_t^* \bar{P}_t^*$$

and

$$(1.26) \quad G_t J_t^* - J_t G_t^* = R_t - R_t^* + \bar{P}_t J_t J_t^* - J_t J_t^* \bar{P}_t^*.$$

Thus, by replacing R_t by \bar{R}_t in these expressions, we deduce approximations for the quadratic variation of H_t and the skew-symmetric part of the quadratic covariation of H_t and Y_t . If we choose $\bar{P}_t = K'I$ for some $K' > 0$, the estimates for $G_t G_t^*$ and $G_t J_t^* - J_t G_t^*$ depend, respectively, only on the symmetric and skew-symmetric parts of $\Gamma_t^* x$, $x \in \mathbb{R}^d$, so that the estimation of these two quantities can be separated. Each of them requires only the computation of Γ_t^E for E symmetric, respectively, for E skew-symmetric; in particular, if we know that $G_t J_t^* = 0$ (independent case), we limit ourselves to E symmetric.

In the case where $J_t = 0$, the above result may still be applied and provides an approximation of the quadratic variation of an observable process H_t ; however, in this case, ε can be chosen arbitrarily, so the precision of the approximation is also arbitrary. This remark can be applied to the estimation of $J_t J_t^*$.

The symmetric part of $G_t J_t^*$ generally cannot be identified, as can be seen, for instance, for linear systems with an unknown parameter. Suppose that W_t is two-dimensional; that H_t is real; that θ is some real random parameter; that the conditional law of H_0 given θ is Gaussian with mean 0, variance $1 - \sin \theta$; and that

$$(1.27) \quad dH_t = \cos \theta dW_t^1 + \sin \theta dW_t^2,$$

$$(1.28) \quad dY_t = H_t dt + \varepsilon dW_t^2.$$

If \mathcal{Y}_t^θ is the filtration generated by θ and Y_t and if \hat{H}_t^θ is the conditional mean of H_t given \mathcal{Y}_t^θ , then \hat{H}_t^θ is given by the Kalman filter

$$(1.29) \quad \hat{H}_0^\theta = 0, \quad d\hat{H}_t^\theta = \frac{1}{\varepsilon} (dY_t - \hat{H}_t^\theta dt).$$

In particular, \hat{H}_t^θ is a Lipschitz function of $(Y_s, s \leq t)$; if we put

$$(1.30) \quad Y_t = \int_0^t \hat{H}_s^\theta ds + \varepsilon \beta_t^\theta,$$

then β_t^θ is a \mathcal{Y}_t^θ Wiener process, and Y_t is adapted to the filtration of β_t^θ . Since β^θ and θ are independent, it appears that Y and θ are independent, so the knowledge of Y does not give any information on θ . In particular, we cannot decide whether the signal and the observation noise are independent ($\theta = 0$). More generally, for the stationary filtering of linear systems with unknown parameter (G, J) such that $JJ^* = I$, parameters yielding the same gain in the Kalman filter are indistinguishable; this is also approximately true for time-varying parameters. To conclude this discussion, we can say that the stationary gain of the Kalman filter of the frozen system $(G, J) = (G_t, J_t)$ is the only quantity that can be identified; in particular, the process R_t estimated in Theorem 2 is a function of this gain. We will now explain how to directly estimate this gain.

1.2. Adaptive filters. Consider again the linear system (1.12) with coefficients (G, J) , suppose that JJ^* is invertible, and let

$$(1.31) \quad S = S(J) = I - J^*(JJ^*)^{-1}J$$

be the orthogonal projection on $\ker J$. Assume that GSG^* is invertible (this condition means that there is some nondegenerate noise in the signal that is independent of the observation noise); then the linear system admits a stationary Kalman filter. Its gain $P = P(G, J)$ is obtained as the solution of $R(G, J, P) = 0$ and is given by

$$(1.32) \quad P = (Q + GJ^*)(JJ^*)^{-1}$$

with $Q = Q(G, J)$ solution of

$$(1.33) \quad Q(JJ^*)^{-1}Q + GJ^*(JJ^*)^{-1}Q + Q(JJ^*)^{-1}JG^* = GSG^*.$$

Let us now return to the time-varying system (0.3), (0.4) and to the filter (1.8), and let us consider which is the best choice of \bar{P}_t for the estimation of H_t . If $P_t = P(G_t, J_t)$ is observable, then we should use $\bar{P}_t = P_t$. However, in our situation, G_t and J_t are not observable, so we want \bar{P}_t to be an approximation of P_t , and, moreover, we want to jointly estimate (H_t, P_t) . To this end, let us take the following viewpoint: We want to make $\int |H_t - \bar{H}_t|^2 dt$ as small as possible, and from (1.11) we have a good observable approximation of its gradient with respect to \bar{P}_t . Thus, we can use a gradient algorithm; for $K > 0$, we let \bar{P}_t be solution of

$$(1.34) \quad d\bar{P}_t = \frac{K}{\varepsilon^{3/2}} \Gamma_t^*(dY_t - \bar{H}_t dt).$$

THEOREM 3. Assume that G_t, J_t , and H_t are of class \mathcal{S} ; that G_t and J_t are bounded; that $J_t J_t^*$ and $G_t S_t G_t^*$ are uniformly elliptic; that $\bar{F}_t = O(1)$, and that $H_0 - \bar{H}_0 = O(\sqrt{\varepsilon})$. Let $P_t = P(G_t, J_t)$ be defined by (1.32) and consider the filter (1.8), (1.10), (1.34) with \bar{P}_0 in a compact subset of M_d^+ . Then

$$(1.35) \quad H_t - \bar{H}_t = O(\sqrt{\varepsilon}) \quad \text{and} \quad P_t - \bar{P}_t = O(\varepsilon^{1/4} + |P_0 - \bar{P}_0|e^{-ct/\sqrt{\varepsilon}})$$

on bounded time intervals, except on an observable event of probability $O(\varepsilon^\infty)$.

Remark 1. The rare event on which we cannot make our estimation is the set where \bar{P}_t becomes too large, too unstable, or has variations that are too fast. Observe indeed that, from our assumptions, P_t stays in a compact subset of M_d^+ ; however, nothing prevents \bar{P}_t from leaving this subset. It may even leave M_d^+ , and, in this case, the system becomes unstable. This can probably be remedied by modifying (1.34) near the boundary of M_d^+ , so that \bar{P}_t is constrained to stay in the above compact subset; after time discretization, the modified filter should reduce to a projected stochastic algorithm as described in [7].

Remark 2. As in § 1.1, we deduce estimates for the quadratic variation and covariation of H and Y via the formulas

$$(1.36) \quad G_t G_t^* = P_t J_t J_t^* P_t^*,$$

$$(1.37) \quad G_t J_t^* - J_t G_t^* = P_t J_t J_t^* - J_t J_t^* P_t^*,$$

which are proved from (1.32) and (1.33).

The filter of Theorem 3 is uniformly stable for the estimation of P_t in the sense that it identifies P_t even if P_0 is badly known. A wider class of filters is obtained by replacing K by a process K_t , taking its values in the space $M_{d \times d}$ of endomorphisms of M_d . If we assume that $P_0 - \bar{P}_0$ is small, we only need local stability on the filter and

obtain the following result: for K in $M_{d \times d}$, U, V in M_d , we let $\rho(K, U, V)$ be the endomorphism of M_d defined by

$$(1.38) \quad \rho(K, U, V): p \mapsto K[UpV].$$

THEOREM 4. *Assume that the conditions of Theorem 3 are satisfied, that $P_0 - \bar{P}_0$ is of order $\varepsilon^{1/4}$, and that K_t is a bounded observable $M_{(d \times d)}$ -valued process such that $K_t - K_s$ is of order $\varepsilon^{1/4}$ for $|t - s| \leq \varepsilon^{3/4}$ (in the sense of (1.2)). Suppose also that, for any bounded symmetric uniformly elliptic processes U_t, V_t of M_d , the process $-\rho(K_t, U_t, V_t)$ is exponentially stable in the timescale $\sqrt{\varepsilon}$. Consider the filter defined by (1.8), (1.10), and*

$$(1.39) \quad d\bar{P}_t = \frac{K_t}{\varepsilon^{3/2}} [\Gamma_t^*(dY_t - \bar{H}_t dt)].$$

Then $H_t - \bar{H}_t$ and $P_t - \bar{P}_t$ are, respectively, of order $\sqrt{\varepsilon}$ and $\varepsilon^{1/4}$ on bounded time intervals, except on an observable event of probability $O(\varepsilon^\infty)$.

Let us now look at the conditional mean \hat{H}_t of H_t given \mathcal{Y}_t . The filters of Theorem 1 satisfy $H_t - \bar{H}_t = O(\sqrt{\varepsilon})$, so, by conditioning on \mathcal{Y}_t , $\hat{H}_t - \bar{H}_t$ is also $O(\sqrt{\varepsilon})$. We now see that adaptive filters are better for the estimation of \hat{H}_t . The following result can be applied to the filters of Theorems 3 and 4.

THEOREM 5. *On a time interval $[T_0, T_1]$, suppose that G_t and J_t are bounded, that G_t, J_t , and H_t are of class \mathcal{F} , that $G_t S_t G_t^*$ is uniformly elliptic, and that $J_t J_t^* = I$. Suppose also that, for each ε fixed, there exists some $\alpha_\varepsilon > 0$ such that*

$$(1.40) \quad \sup_{T_0 \leq t \leq T_1} \mathbb{E} \exp \alpha_\varepsilon |F_t|^2 < \infty.$$

Let \bar{F}_t and \bar{P}_t be observable processes and consider the associate filter (1.8). We suppose that $\bar{F}_t = O(1)$ and that, except on an event of probability $O(\varepsilon^\infty)$, $H_t - \bar{H}_t$ and $P_t - \bar{P}_t$ are, respectively, $O(\sqrt{\varepsilon})$ and $O(\varepsilon^{1/4})$ on $[T_0, T_1]$. Then, for any $T'_0 \in (T_0, T_1)$, the process $\bar{H}_t - \hat{H}_t$ is of order $\varepsilon^{3/4}$ in probability on the time interval $[T'_0, T_1]$.

Remark. This result is false without the assumption on $J_t J_t^*$; if indeed $J_t J_t^*$ is random, then H_t may be observable, for instance, when H_t is a function of $J_t J_t^*$. In such a case, \hat{H}_t is equal to H_t , but $H_t - \bar{H}_t$ is not of order $\varepsilon^{3/4}$.

The next step would consist in giving a more precise estimation of $P_t - \bar{P}_t$ and in comparing the efficiency of the filters obtained for various gains K_t .

1.3. Filtering of Markov diffusion processes. In the Markovian case described in the Introduction, the processes P_t, Q_t , and S_t defined by (1.31)–(1.33), respectively, are equal to $p(X_t), q(X_t)$, and $\sigma(X_t)$, where

$$(1.41) \quad p = (q + h' g j^*)(j j^*)^{-1}, \quad \sigma = I - j^*(j j^*)^{-1} j,$$

$$q(j j^*)^{-1} q + h' g j^*(j j^*)^{-1} q + q(j j^*)^{-1} j g^* h'^* = h' g \sigma g^* h'^*.$$

We also use the function p^\dagger obtained from p by (1.4). We follow previous results, but we also must consider the geometric structure of the process (X_t, Y_t) . We are only interested here in the local tracking of X_t , so we suppose that $X_0 - \bar{X}_0$ is small.

THEOREM 6. *Let f, g, j , and h be fixed Borel functions defined on \mathbb{R}^n . Assume that f has at most linear growth at infinity; that g, j , and h' are C_b^2 ; that $j j^*$ and $h' g \sigma g^* h'^*$ are uniformly elliptic; that X_0 is bounded; and that (X_t, Y_t) is the solution of (0.5), (0.6). Let ϕ and ψ be fixed C_b^2 and C_b^1 functions, respectively, defined on \mathbb{R}^n such that, for each x , $\phi(x)$ is a linear map from \mathbb{R}^d into \mathbb{R}^n , $\psi(x)$ is a linear map from M_d into \mathbb{R}^n ; we suppose that $h' \phi = p$, that $h' \psi = 0$ (so that the image of $\psi(x)$ lies in $\ker h'(x)$), and that the real parts of the eigenvalues of the maps*

$$(1.42) \quad \rho(x): \ker h'(x) \rightarrow \ker h'(x), \quad \mu \mapsto \psi(x)[p^\dagger(x) \nabla_\mu p(x) j j^*(x)]$$

are bounded below by a positive constant. Let \bar{X}_t be the solution of

$$(1.43) \quad d\bar{X}_t = \frac{1}{\varepsilon} \phi(\bar{X}_t)(dY_t - h(\bar{X}_t) dt) + \frac{1}{\varepsilon^{3/2}} \psi(\bar{X}_t)[\Gamma_t^*(dY_t - h(\bar{X}_t) dt)],$$

$$(1.44) \quad d\Gamma_t^E = -\frac{1}{\varepsilon} p(\bar{X}_t) \Gamma_t^E dt + \frac{E}{\varepsilon} (dY_t - h(\bar{X}_t) dt).$$

Suppose that $X_0 - \bar{X}_0$ is of order $\varepsilon^{1/4}$ and that $h(X_0) - h(\bar{X}_0)$ is of order $\sqrt{\varepsilon}$. Then

$$(1.45) \quad X_t - \bar{X}_t = O(\varepsilon^{1/4}), \quad h(X_t) - h(\bar{X}_t) = O(\sqrt{\varepsilon})$$

on bounded time intervals, except on an event of probability $O(\varepsilon^\infty)$. Moreover, if $jj^* = I$ and if \hat{H}_t is the conditional expectation of $h(X_t)$ given \mathcal{Y}_t , then $\hat{H}_t - h(\bar{X}_t)$ is of order $\varepsilon^{3/4}$ in probability on any fixed time interval $[T_0, T_1] \subset (0, \infty)$.

Remark 3. We do not claim in this result that the rare event on which we cannot make our estimation is observable.

Remark 4. The ellipticity and boundedness assumptions imply that p takes its values in a compact subset of M_d^+ , so p^\dagger is bounded and elliptic. If $gj^* = 0$ (independent noise) and $jj^* = I$, then $p = (h'gg^*h'^*)^{1/2}$ is symmetric and $p^\dagger = p^{-1}$.

Remark 5. As in Theorem 5, the estimation of \hat{H}_t requires an additional assumption on jj^* ; it is indeed well known (see [17]) that the optimal filter is singular if $jj^*(x)$ depends on x . Without the assumption $jj^* = I$, we can only say that $h(\bar{X}_t)$ is the best approximation of $h(X_t)$ among filters of type (1.43).

Remark 6. Our conditions are local, so we must assume that $X_0 - \bar{X}_0$ is small. To drop this condition, we may conjecture that the stability condition should be the following one: For any fixed x and any \bar{x}_0 in the manifold $\{h(\bar{x}) = h(x)\}$, the solution of

$$(1.46) \quad \dot{\bar{x}}_t = -\psi(\bar{x}_t) \nabla \Lambda(g(x), j(x), p(\bar{x}_t))$$

converges to x as $t \rightarrow \infty$. Roughly speaking, the condition about the eigenvalues of $\rho(x)$ means that this holds for \bar{x}_0 in a neighbourhood of x .

Remark 7. As before, the next step would consist in estimating more precisely $X_t - \bar{X}_t$ to compare the efficiency of the filters corresponding to various ϕ and ψ . In the particular example of [15], a precise estimation of \hat{X}_t was derived; we found a filter \bar{X}_t such that $\hat{X}_t - \bar{X}_t$ and $\hat{H}_t - h(\bar{X}_t)$ are, respectively, $O(\sqrt{\varepsilon})$ and $O(\varepsilon)$. However, this example is particular in the sense that it is partly linear, and we do not know how to prove similar properties in the general case.

Example. Let us describe a filter satisfying the conditions of the theorem. It follows from the assumptions on the coefficients that $h'h'^*$ is invertible, so we can use

$$(1.47) \quad \phi = h'^*(h'h'^*)^{-1}p.$$

We can also let $\psi(x)$ be defined by its adjoint

$$(1.48) \quad \ker h'(x) \rightarrow M_d, \quad \mu \mapsto p^\dagger(x) \nabla_\mu p(x).$$

Then, for μ and ν in $\ker h'(x)$, we have

$$\begin{aligned} (\rho(x)\mu)^*\nu &= \text{trace}((\rho(x)\mu)\nu^*) \\ (1.49) \quad &= \text{trace}(\psi(x)[p^\dagger(x)\nabla_\mu p(x)jj^*(x)]\nu^*) \\ &= \text{trace}(p^\dagger(x)\nabla_\mu p(x)jj^*(x)\nabla_\nu p(x)^*p^\dagger(x)^*). \end{aligned}$$

In particular, $\rho(x)$ is symmetric, so the assumption on its eigenvalues is satisfied as soon as it is uniformly elliptic; this holds under the condition

$$(1.50) \quad |\nabla_{\mu} p(x)| \geq c|\mu|$$

for any μ in $\ker h'(x)$. This condition means that $x \mapsto (h(x), p(x))$ is locally injective.

2. Proofs of Theorems 1 and 2.

2.1. Exponentially stable systems and Theorem 1. We first summarize the basic properties of exponentially stable processes, which will be used later. Some of these properties were proved in [14]; see also [4].

LEMMA 1. *Let A_t be a bounded family of matrix-valued adapted processes. Then*

(1) *The three following conditions are equivalent:*

- (i) *The process A_t is exponentially stable in the timescale ε^{α} ;*
- (ii) *The solution λ_t of*

$$(2.1) \quad \varepsilon^{\alpha} \dot{\lambda}_t = A_t \lambda_t + \lambda_t A_t^* + I, \quad \lambda_0 = I$$

is bounded;

- (iii) *There exists an absolutely continuous process λ_t with values in symmetric matrices, such that $\lambda_0 \geq I$, λ_t is bounded and*

$$(2.2) \quad \varepsilon^{\alpha} \dot{\lambda}_t - A_t \lambda_t - \lambda_t A_t^* \geq I.$$

(2) *If A_t is exponentially stable in some timescale, then there exists a constant $C > 0$ such that any \bar{A}_t satisfying $|\bar{A}_t - A_t| \leq C$ is exponentially stable in the same timescale.*

(3) *Suppose that the real parts of the eigenvalues of A_t are bounded above by a negative constant, and that, for $|t - s| \leq \varepsilon^{\alpha}$, we have*

$$(2.3) \quad |A_t - A_s| \leq C\varepsilon^{\beta}$$

for some positive β and C ; then A_t is exponentially stable in the timescale ε^{α} .

(4) *Suppose that A_t is exponentially stable in the timescale ε^{α} , that Z_t satisfies*

$$(2.4) \quad dZ_t = \frac{1}{\varepsilon^{\alpha}} A_t Z_t dt + \frac{1}{\varepsilon^{\alpha}} f_t dt + \frac{1}{\varepsilon^{\alpha/2}} g_t dW_t,$$

and that f_t and g_t are of order ε^{γ} uniformly for $t \geq 0$. Then

$$(2.5) \quad Z_t = O(\varepsilon^{\gamma} + |Z_0|e^{-ct/\varepsilon^{\alpha}})$$

for some $c > 0$, uniformly for $t \geq 0$.

Proof of part (1). To prove that (i) implies (ii), note that, if Z_t is solution of (1.6)

$$(2.6) \quad \lambda_t = Z_t Z_t^* + \varepsilon^{-\alpha} \int_0^t Z_s Z_s^{-1} (Z_s Z_s^{-1})^* ds$$

is bounded from (1.7). It is immediate that (ii) implies (iii). To prove that (iii) implies (i), we look at the equation satisfied by $(Z_t Z_s^{-1})^* \lambda_t^{-1} Z_t Z_s^{-1}$ for $t \geq s$, and, by means of the inequality $\lambda_t^{-2} \geq \lambda_t^{-1}/|\lambda_t|$, we deduce that it decreases exponentially fast.

Proof of part (2). In condition (iii), it is clear that, by multiplying λ_t by a positive constant, we can replace the condition “ $\geq I$ ” by the condition “uniformly elliptic.” Since λ_t is bounded, this property also holds for \bar{A}_t in a neighbourhood of A_t .

Proof of part (3). Consider the time subdivision $s_i = i\varepsilon^{\alpha}$; define λ_{s_i} by

$$(2.7) \quad A_{s_i} \lambda_{s_i} + \lambda_{s_i} A_{s_i}^* + \gamma I = 0,$$

where $\gamma > 1$ is such that $\lambda_0 \geq I$; and let λ_t be the affine interpolation of these values. Then λ_t is bounded; moreover, for $s_i < t < s_{i+1}$, we have

$$(2.8) \quad |\lambda_t - \lambda_{s_i}| \leq C\varepsilon^\beta, \quad |\dot{\lambda}_t| \leq C\varepsilon^{\beta-\alpha},$$

so

$$(2.9) \quad \varepsilon^\alpha \dot{\lambda}_t - A_t \lambda_t - \lambda_t A_t^* \geq (\gamma - C\varepsilon^\beta) I \geq I$$

for ε small enough. Therefore, the exponential stability of A_t is deduced from the characterization (iii).

Sketch of the proof of part (4). We do not give the whole proof of this result, which is easily deduced from Lemma 1.3.2 of [14]. We take a process λ_t satisfying (iii), and, by applying Ito's formula, we write the differential increment of $Z_t^* \lambda_t^{-1} Z_t$. From (iii), we deduce that the moments of this process satisfy differential inequations, and therefore, we obtain estimates on these moments. \square

Remark. By considering the trace of λ_t^{-1} , it can be checked that a process λ_t satisfying (2.2) is uniformly elliptic as soon as A_t is bounded.

Proof of Theorem 1. By putting $Z_t = H_t - \bar{H}_t$, we have

$$(2.10) \quad dZ_t = -\frac{\bar{P}_t}{\varepsilon} Z_t dt + (F_t - \bar{F}_t) dt + (G_t - \bar{P}_t J_t) dW_t,$$

so we must only apply part (4) of Lemma 1. \square

2.2. An averaging principle. Assume the conditions of Theorem 2. Observe that part (4) of Lemma 1 implies that Γ_t is of order $\sqrt{\varepsilon}$. Since the estimation problem involves two timescales, the idea is to use a discretization of the time with a meshsize intermediate between the two timescales, more precisely, $\varepsilon^{3/4}$. Thus we define $t_i = i\varepsilon^{3/4}$. The notation $O_i(\varepsilon^\gamma)$ will mean of order ε^γ with zero conditional mean given \mathcal{F}_{t_i} .

Lemma 2, below, is an averaging principle; it will be basic in the proof of all our results. In classical results such as [8], we consider a fast ergodic process ξ_t and a slow process, which is solution of a stochastic differential equation, the coefficients of which depend on ξ_t . Then we prove that the behaviour of the slow component can be approximated by taking the average of the coefficients with respect to the invariant measure of ξ_t . In the proof of the following result, the statistics of the fast component are allowed to vary slowly.

LEMMA 2. *In the assumptions of Theorem 2, relax condition (1.23) on the oscillations of \bar{P} in the following one. We suppose that $\bar{P}_t - \bar{P}_s$ is of order $\varepsilon^{1/4}$ for $|t - s| \leq \varepsilon^{3/4}$. Then, for $i \geq 0$, we have*

$$(2.11) \quad \varepsilon^{-7/4} \int_{t_i}^{t_{i+1}} \Gamma_u^* (dY_u - \bar{H}_u du) = \frac{1}{2} \bar{P}_{t_i}^* R_{t_i} + O_i(\varepsilon^{1/8}) + O(\varepsilon^{1/4}).$$

Proof. By developing dY_u , observe that

$$(2.12) \quad \varepsilon^{-7/4} \int_{t_i}^{t_{i+1}} \Gamma_u^* (dY_u - \bar{H}_u du) = \varepsilon^{-7/4} \int_{t_i}^{t_{i+1}} \Gamma_u^* (H_u - \bar{H}_u) du + O_i(\varepsilon^{1/8}).$$

Moreover, since Γ_u^* is the adjoint of Γ_u , for any E in M_d , we have

$$(2.13) \quad \text{trace} \left(\int_{t_i}^{t_{i+1}} \Gamma_u^* (H_u - \bar{H}_u) du E^* \right) = \int_{t_i}^{t_{i+1}} (H_u - \bar{H}_u)^* \Gamma_u^E du,$$

so it is sufficient to estimate this quantity for E fixed and prove that

$$(2.14) \quad \varepsilon^{-7/4} \int_{t_i}^{t_{i+1}} (H_u - \bar{H}_u)^* \Gamma_u^E du = \frac{1}{2} \text{trace} (\bar{P}_{t_i}^* R_{t_i} E^*) + O_i(\varepsilon^{1/8}) + O(\varepsilon^{1/4}).$$

For $t_i \leq t < t_{i+1}$, consider the process (η_t, ξ_t) , which is the solution of

$$(2.15) \quad d\eta_t = -\frac{\bar{P}_{t_i}}{\varepsilon} \eta_t dt + \frac{G_{t_i} - \bar{P}_{t_i} J_{t_i}}{\sqrt{\varepsilon}} dW_t,$$

$$(2.16) \quad d\xi_t = -\frac{\bar{P}_{t_i}}{\varepsilon} \xi_t dt + \frac{E}{\varepsilon} \eta_t dt + \frac{E J_{t_i}}{\sqrt{\varepsilon}} dW_t,$$

with $\eta_t = (H_t - \bar{H}_{t_i})/\sqrt{\varepsilon}$ and $\xi_t = \Gamma_{t_i}^E/\sqrt{\varepsilon}$; observe that $\xi_t = -\nabla_E \eta_t$. The variables \bar{P}_{t_i} are in a compact subset of M_d^+ , so we can deduce from part (4) of Lemma 1 that η_t and ξ_t are $O(1)$. We have assumed that G_t and J_t are of class \mathcal{J} , so their oscillations on $[t_i, t_{i+1}]$ are of order $\varepsilon^{3/8}$; we have also assumed that the oscillations of \bar{P}_t are of order $\varepsilon^{1/4}$. Thus, by comparing the above equations with the equations of $H_t - \bar{H}_{t_i}$ and Γ_t^E and by applying part (4) of Lemma 1, we obtain that

$$(2.17) \quad \eta_t = \frac{H_t - \bar{H}_{t_i}}{\sqrt{\varepsilon}} + O(\varepsilon^{1/4}) \quad \text{and} \quad \xi_t = \frac{\Gamma_t^E}{\sqrt{\varepsilon}} + O(\varepsilon^{1/4}).$$

Conditionally on \mathcal{F}_{t_i} , the process (η_t, ξ_t) , $t_i \leq t < t_{i+1}$ is Gaussian; moreover, since \bar{P}_{t_i} is in a compact subset of M_d^+ , for $t_i \leq s \leq t < t_{i+1}$, the conditional covariance of (η_s, ξ_s) and (η_t, ξ_t) is of order $\exp(-c(t-s)/\varepsilon)$. In particular, the conditional covariance of $\eta_s^* \xi_s$ and $\eta_t^* \xi_t$ satisfies an estimate of the same type, so the conditional variance of $\varepsilon^{-3/4} \int_{t_i}^{t_{i+1}} \eta_u^* \xi_u du$ is of order $\varepsilon^{1/4}$. Thus

$$(2.18) \quad \begin{aligned} \varepsilon^{-7/4} \int_{t_i}^{t_{i+1}} (H_u - \bar{H}_{t_i})^* \Gamma_u^E du &= \varepsilon^{-3/4} \int_{t_i}^{t_{i+1}} \eta_u^* \xi_u du + O(\varepsilon^{1/4}) \\ &= \varepsilon^{-3/4} \int_{t_i}^{t_{i+1}} \mathbb{E}[\eta_u^* \xi_u | \mathcal{F}_{t_i}] du + O_i(\varepsilon^{1/8}) + O(\varepsilon^{1/4}). \end{aligned}$$

Then, for $t_i \leq t < t_{i+1}$, we write the equations satisfied by $\eta_t \eta_t^*$ and $\eta_t \xi_t^*$, and we compute their conditional expectations. We deduce that with an error of order $\exp(-c(t-t_i)/\varepsilon)$, the conditional expectation of $\eta_t \eta_t^*$ is the asymptotic covariance error for the linear system with coefficients (G_{t_i}, J_{t_i}) and gain \bar{P}_{t_i} . Thus

$$(2.19) \quad \mathbb{E}[\eta_t \eta_t^* | \mathcal{F}_{t_i}] = \bar{Q}(G_{t_i}, J_{t_i}, \bar{P}_{t_i}) + O(e^{-c(t-t_i)/\varepsilon}).$$

Similarly, for $\eta_t \xi_t^*$, we check that we can differentiate the previous estimate, and we obtain

$$(2.20) \quad \mathbb{E}[\eta_t \xi_t^* + \xi_t \eta_t^* | \mathcal{F}_{t_i}] = -\nabla_E \bar{Q}(G_{t_i}, J_{t_i}, \bar{P}_{t_i}) + O(e^{-c(t-t_i)/\varepsilon}).$$

By taking the trace and using (1.16),

$$(2.21) \quad \begin{aligned} \mathbb{E}[\eta_t^* \xi_t | \mathcal{F}_{t_i}] &= -\frac{1}{2} \nabla_E \Lambda(G_{t_i}, J_{t_i}, \bar{P}_{t_i}) + O(e^{-c(t-t_i)/\varepsilon}) \\ &= \frac{1}{2} \text{trace}(\bar{P}_{t_i}^* R_{t_i} E^*) + O(e^{-c(t-t_i)/\varepsilon}), \end{aligned}$$

so, by integrating on $[t_i, t_{i+1}]$ and applying (2.18), estimate (2.14), which was required, is proved. \square

2.3. Linear stochastic algorithms and Theorem 2.

LEMMA 3. Let Z_t be a nonnegative adapted step process for the subdivision $t_i = i\varepsilon^{3/4}$. Suppose that $Z_0 = 0$ and that

$$(2.22) \quad Z_{t_{i+1}} \leq Z_{t_i} - c\varepsilon^{1/4} Z_{t_i} + (1 + \sqrt{Z_{t_i}})(O_i(\varepsilon^{1/8}) + O(\varepsilon^{1/4}))$$

for some $c > 0$. Then Z_t is $O(1)$.

Proof. Let $k \geq 2$ be an integer. From the inequality

$$(2.23) \quad |(a+b)^k - a^k - ka^{k-1}b| \leq Ca^{k-2}b^2 + Cb^k$$

applied with $a+b$, the right-hand side of (2.22), and $a = Z_{t_i}(1 - c\varepsilon^{1/4})$, we obtain

$$(2.34) \quad \begin{aligned} Z_{t_{i+1}}^k &\leq Z_{t_i}^k(1 - c\varepsilon^{1/4})^k + (Z_{t_i}^{k-1/2} + Z_{t_i}^{k-1})O_i(\varepsilon^{1/8}) \\ &\quad + (Z_{t_i}^{k-1/2} + Z_{t_i}^{k-2})O(\varepsilon^{1/4}) + (1 + Z_{t_i}^{k/2})O(\varepsilon^{k/8}). \end{aligned}$$

By taking the expectation,

$$(2.25) \quad \begin{aligned} \mathbb{E}[Z_{t_{i+1}}^k] &\leq \mathbb{E}[Z_{t_i}^k](1 - c\varepsilon^{1/4})^k + \mathbb{E}[Z_{t_i}^{k-1/2}O(\varepsilon^{1/4})] \\ &\quad + \mathbb{E}[Z_{t_i}^{k-2}O(\varepsilon^{1/4})] + \mathbb{E}[Z_{t_i}^{k/2}O(\varepsilon^{k/8})] + O(\varepsilon^{k/8}). \end{aligned}$$

Thus, if we let u_i be the k th moment of Z_{t_i} , by using Hölder's inequality in the last three expectations of previous formula and the estimate

$$(2.26) \quad (1 - c\varepsilon^{1/4})^k = 1 - ck\varepsilon^{1/4} + O(\sqrt{\varepsilon}),$$

we obtain

$$(2.27) \quad \begin{aligned} u_{i+1} &= u_i - ck\varepsilon^{1/4}u_i + u_iO(\sqrt{\varepsilon}) + u_i^{(2k-1)/(2k)}O(\varepsilon^{1/4}) \\ &\quad + u_i^{(k-2)/k}O(\varepsilon^{1/4}) + (1 + \sqrt{u_i})O(\varepsilon^{k/8}). \end{aligned}$$

There is some constant $\gamma > 0$ such that, for ε small enough, $u_{i+1} - u_i \leq 0$ for $u_i \geq \gamma$; moreover, for $u_i \leq \gamma$, $u_{i+1} - u_i$ is of order $\varepsilon^{1/4}$. Thus

$$(2.28) \quad u_i \leq \gamma + O(\varepsilon^{1/4}) = O(1). \quad \square$$

LEMMA 4. Let A_t be a bounded adapted matrix-valued process that is exponentially stable in the timescale $\sqrt{\varepsilon}$ and such that $A_t - A_s$ is of order $\varepsilon^{1/4}$ for $|t - s| \leq \varepsilon^{3/4}$. Let Z_t be an adapted vector-valued step process such that $Z_t = O(1)$ and

$$(2.29) \quad Z_{t_{i+1}} = Z_{t_i} + \varepsilon^{1/4}A_{t_i}Z_{t_i} + O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}).$$

Let z_t be the solution of

$$(2.30) \quad \dot{z}_t = \frac{1}{\sqrt{\varepsilon}}A_t z_t, \quad z_0 = Z_0.$$

Then $Z_t - z_t$ is of order $\varepsilon^{1/4}$; in particular,

$$(2.31) \quad Z_t = O(\varepsilon^{1/4} + |Z_0|e^{-ct/\sqrt{\varepsilon}}).$$

Proof. This lemma can be viewed as a discrete-time analogue of part (4) of Lemma 1; the recursive expression (2.29) can be interpreted as a stochastic algorithm, and (2.30) is the limiting ordinary differential equation. Since z_t is of order $|Z_0|e^{-ct/\sqrt{\varepsilon}}$, note that we must only prove the estimate on $Z_t - z_t$, and (2.31) will follow; moreover, from our assumption on the oscillations of A_t ,

$$(2.32) \quad z_{t_{i+1}} = z_{t_i} + \varepsilon^{1/4}A_{t_i}z_{t_i} + O(\sqrt{\varepsilon}).$$

Thus Z_t and $Z_t - z_t$ satisfy the same recursive estimate (2.29), and the proof can be reduced to the case where $Z_0 = 0$. Let λ_t be a process satisfying the conditions of the characterization (iii) of Lemma 1 and put $V_t = \lambda_t^{-1}$, so that V_t is bounded, uniformly elliptic, and

$$(2.33) \quad \sqrt{\varepsilon} \dot{V}_t + V_t A_t + A_t^* V_t \leq -cV_t.$$

By integrating this inequality,

$$(2.34) \quad \varepsilon^{-1/4}(V_{t_{i+1}} - V_{t_i}) + V_{t_i}A_{t_i} + A_{t_i}^*V_{t_i} \leq -cV_{t_i} + O(\varepsilon^{1/4}).$$

On the other hand, by writing the recursive formula satisfied by $Z_{t_i}^* V_{t_i} Z_{t_i}$ and by applying the recursive estimate (2.29) of Z_{t_i} , we obtain

$$(2.35) \quad \begin{aligned} Z_{t_{i+1}}^* V_{t_{i+1}} Z_{t_{i+1}} &= Z_{t_i}^* V_{t_i} Z_{t_i} + Z_{t_i}^* (V_{t_{i+1}} - V_{t_i} + \varepsilon^{1/4} A_{t_i}^* V_{t_i} + \varepsilon^{1/4} V_{t_i} A_{t_i}) Z_{t_i} \\ &\quad + |Z_{t_i}|^2 O(\sqrt{\varepsilon}) + |Z_{t_i}| (O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon})) + O(\varepsilon^{3/4}). \end{aligned}$$

We have assumed that $Z_t = O(1)$, so

$$(2.36) \quad |Z_{t_i}|^2 O(\sqrt{\varepsilon}) = |Z_{t_i}| O(\sqrt{\varepsilon}).$$

By also applying (2.34), we obtain

$$(2.37) \quad \begin{aligned} Z_{t_{i+1}}^* V_{t_{i+1}} Z_{t_{i+1}} &\leq Z_{t_i}^* V_{t_i} Z_{t_i} (1 - c\varepsilon^{1/4}) + |Z_{t_i}| (O_i(\varepsilon^{3/8}) \\ &\quad + O(\sqrt{\varepsilon})) + O(\varepsilon^{3/4}). \end{aligned}$$

Thus $Z_t^* V_t Z_t / \sqrt{\varepsilon}$ satisfies the assumptions of Lemma 3, and so is $O(1)$. \square

Proof of Theorem 2. We have

$$(2.38) \quad d\bar{R}_t = \frac{K}{\sqrt{\varepsilon}} \left(O(1) dt + O(\sqrt{\varepsilon}) dW_t - \frac{1}{2} \bar{P}_t^+ \bar{R}_t dt \right).$$

On the other hand, since \bar{P}_t lies in a compact subset of M_d^+ , we deduce that \bar{P}_t^+ is bounded and uniformly elliptic; thus it is exponentially stable in all timescales, and we deduce from part (4) of Lemma 1 that $\bar{R}_t = O(1)$. By again applying (2.38), the oscillations of \bar{R}_t on intervals of length $\varepsilon^{3/4}$ are shown to be of order $\varepsilon^{1/4}$; \bar{P}_t^+ satisfies the same condition, so

$$(2.39) \quad \bar{R}_{t_{i+1}} - \bar{R}_{t_i} = \frac{K}{\varepsilon^{3/2}} \int_{t_i}^{t_{i+1}} \Gamma_t^*(dY_t - \bar{H}_t dt) - \frac{K}{2} \varepsilon^{1/4} \bar{P}_{t_i}^+ \bar{R}_{t_i} + O(\sqrt{\varepsilon})$$

and, from Lemma 2,

$$(2.40) \quad \bar{R}_{t_{i+1}} - \bar{R}_{t_i} = \frac{K}{2} \varepsilon^{1/4} \bar{P}_{t_i}^+ (R_{t_i} - \bar{R}_{t_i}) + O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}).$$

On the other hand, $R_t = R(G_t, J_t, \bar{P}_t)$, and R is a smooth function; since G_t and J_t are of class \mathcal{F} and from assumption (1.23) about the oscillations of \bar{P}_t , we have

$$(2.41) \quad R_{t_{i+1}} - R_{t_i} = O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}).$$

Thus the step process Z_t defined by $Z_{t_i} = R_{t_i} - \bar{R}_{t_i}$ satisfies the assumptions of Lemma 4 and therefore is of order $\varepsilon^{1/4} + |Z_0| e^{-ct/\sqrt{\varepsilon}}$. \square

3. Proofs of Theorems 3–5.

3.1. Nonlinear stochastic algorithms.

LEMMA 5. *Let Z_t be a bounded adapted step process associated to the subdivision (t_i) , with values in a Euclidean space E . We suppose that*

$$(3.1) \quad Z_{t_{i+1}} = Z_{t_i} + \varepsilon^{1/4} b(\beta_{t_i}, Z_{t_i}) + O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}),$$

where β_t is a process of class \mathcal{F} taking its values in a compact subset \mathcal{H}_0 of a Euclidean space, and b is a smooth function on $\mathcal{H}_0 \times E$; we denote $b' = \partial b / \partial z$. We also suppose that $b(\beta, 0) = 0$ for any β , that the real parts of the eigenvalues of $b'(\beta, 0)$ are bounded above by a negative constant, that $b'(\beta_t, 0)$ is exponentially stable in the timescale $\sqrt{\varepsilon}$, and that, for any β and z_0 , the solution of

$$(3.2) \quad \dot{z}_t^\beta = b(\beta, z_t^\beta), \quad z_0^\beta = z_0$$

converges to zero as $t \rightarrow +\infty$. Let z_t be the solution of

$$(3.3) \quad \dot{z}_t = \frac{1}{\sqrt{\varepsilon}} b(\beta_0, z_t), \quad z_0 = Z_0.$$

Then $Z_t - z_t$ is of order $\varepsilon^{1/4}$; in particular, Z_t is of order $\varepsilon^{1/4} + |Z_0|e^{-ct/\sqrt{\varepsilon}}$.

Remark 8. This result can be viewed as a nonlinear version of Lemma 4. The idea of the proof is to transform the differential equation into an almost linear one, so that we can apply Lemma 4.

Remark 9. The assumption about the solution of (3.2) implies that the real parts of the eigenvalues of $b'(\beta, 0)$ are nonpositive; thus the additional assumption about these eigenvalues simply means that they should not be on the imaginary axis.

Proof. From part (2) of Lemma 1, since $b'(\beta_t, 0)$ and $b'(\beta_0, 0)$ are bounded and exponentially stable, there exists a $\lambda > 0$ such that any A_t satisfying

$$(3.4) \quad |A_t - b'(\beta_t, 0)| \leq \lambda \quad \text{or} \quad |A_t - b'(\beta_0, 0)| \leq \lambda$$

is exponentially stable. On the other hand, there exists a matrix-valued smooth function a such that $b(\beta, z) = a(\beta, z)z$; let $\tilde{a}(\beta, z)$ and $\tilde{b}(\beta, z) = \tilde{a}(\beta, z)z$ be other smooth functions that coincide with a and b for z in a neighbourhood of zero such that

$$(3.5) \quad |\tilde{a}(\beta, z) - b'(\beta, 0)| \leq \lambda$$

and such that, for any β and \tilde{z}_0 , the solution of

$$(3.6) \quad \frac{d}{dt} \tilde{z}_t^\beta = \tilde{b}(\beta, \tilde{z}_t^\beta), \quad \tilde{z}_0^\beta = \tilde{z}_0$$

converges to zero as $t \rightarrow +\infty$. Then $\tilde{a}(\beta_t, e_t)$ and $\tilde{a}(\beta_0, e_t)$ are exponentially stable for any process e_t . Since b and \tilde{b} coincide on a neighbourhood of zero, for any β , there exists a smooth diffeomorphism $D(\beta, \cdot)$, which transforms the flow Φ_β of (3.2) into the flow $\tilde{\Phi}_\beta$ of (3.6); it is given by

$$(3.7) \quad D(\beta, z) = \tilde{\Phi}_\beta(-t, \Phi_\beta(t, z))$$

for t sufficiently large, so that $(\Phi_\beta(s, z), s \geq t)$ stays in the above neighbourhood of zero. Similarly, the inverse of D is given by

$$(3.8) \quad D^{-1}(\beta, z) = \Phi_\beta(-t, \tilde{\Phi}_\beta(t, z))$$

for t large enough. Moreover, D and D^{-1} are smooth. The property of transformation of flows can be translated as

$$(3.9) \quad \frac{\partial D}{\partial z}(\beta, z)b(\beta, z) = \tilde{b}(\beta, D(\beta, z)).$$

Put $\tilde{Z}_t = D(\beta_t, Z_t)$; then

$$(3.10) \quad \begin{aligned} \tilde{Z}_{t_{i+1}} = \tilde{Z}_t &+ \frac{\partial D}{\partial z}(\beta_t, Z_t)(Z_{t_{i+1}} - Z_t) + \frac{\partial D}{\partial \beta}(\beta_t, Z_t)(\beta_{t_{i+1}} - \beta_t) \\ &+ O(|Z_{t_{i+1}} - Z_t|^2) + O(|\beta_{t_{i+1}} - \beta_t|^2). \end{aligned}$$

By using the recursive expression (3.1) for Z_t and by applying (3.9), we obtain

$$(3.11) \quad \tilde{Z}_{t_{i+1}} = \tilde{Z}_t + \varepsilon^{1/4} \tilde{b}(\beta_t, \tilde{Z}_t) + O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}).$$

The assumptions of Lemma 4 are satisfied with $A_t = \tilde{a}(\beta_t, \tilde{Z}_t)$; thus, if ζ_t is the solution of

$$(3.12) \quad \dot{\zeta}_t = \frac{1}{\sqrt{\varepsilon}} \tilde{a}(\beta_t, \tilde{Z}_t) \zeta_t, \quad \zeta_0 = \tilde{Z}_0,$$

then $\zeta_t - \tilde{Z}_t$ is of order $\varepsilon^{1/4}$. By using this estimate and

$$(3.13) \quad |\beta_t - \beta_0| \zeta_t = O(\sqrt{t} e^{-ct/\sqrt{\varepsilon}}) = O(\varepsilon^{1/4}),$$

we deduce that

$$(3.14) \quad \dot{\zeta}_t = \frac{1}{\sqrt{\varepsilon}} (\tilde{a}(\beta_0, \zeta_t) \zeta_t + O(\varepsilon^{1/4})).$$

Moreover, if \tilde{z}_t is the solution of

$$(3.15) \quad \frac{d}{dt} \tilde{z}_t = \frac{1}{\sqrt{\varepsilon}} \tilde{b}(\beta_0, \tilde{z}_t), \quad \tilde{z}_0 = \tilde{Z}_0,$$

then

$$(3.16) \quad \frac{d}{dt} \tilde{z}_t = \frac{1}{\sqrt{\varepsilon}} (\tilde{a}(\beta_0, \zeta_t) \tilde{z}_t + \gamma_t (\tilde{z}_t - \zeta_t)),$$

where γ_t is a process satisfying

$$(3.17) \quad |\gamma_t| \leq C |\tilde{z}_t| \leq C e^{-ct/\sqrt{\varepsilon}}.$$

Thus

$$(3.18) \quad \frac{d}{dt} (\zeta_t - \tilde{z}_t) = \frac{1}{\sqrt{\varepsilon}} (\tilde{a}(\beta_0, \zeta_t) + \gamma_t) (\zeta_t - \tilde{z}_t) + \frac{1}{\sqrt{\varepsilon}} O(\varepsilon^{1/4}).$$

The process $\tilde{a}(\beta_0, \zeta_t)$ is exponentially stable in the timescale $\sqrt{\varepsilon}$, and, from (3.17), a perturbation by γ_t does not destroy this stability. So $\zeta_t - \tilde{z}_t$ is $O(\varepsilon^{1/4})$, and therefore $\tilde{Z}_t - \tilde{z}_t$ is also $O(\varepsilon^{1/4})$. Now we apply

$$(3.19) \quad Z_t = D^{-1}(\beta_t, \tilde{Z}_t), \quad z_t = D^{-1}(\beta_0, \tilde{z}_t).$$

The function $D^{-1}(\beta, \cdot)$ is the identity on a neighbourhood of zero; if C is chosen large enough, for $t \geq C\sqrt{\varepsilon}$, \tilde{z}_t in this neighbourhood, so z_t is equal to $D^{-1}(\beta_t, \tilde{z}_t)$, and the Lipschitz property of D^{-1} with respect to z implies that $Z_t - z_t$ is $O(\varepsilon^{1/4})$. For $t < C\sqrt{\varepsilon}$, we use

$$(3.20) \quad |Z_t - z_t| \leq C |\beta_t - \beta_0| + C |\tilde{Z}_t - \tilde{z}_t|$$

to obtain the same estimate. \square

3.2. Uniformly stable filters and Theorem 3. Assume the conditions of Theorem 3; we first describe the observable event of probability $1 - O(\varepsilon^\infty)$ on which we will estimate $H_t - \bar{H}_t$ and $P_t - \bar{P}_t$. We have supposed that (G_t, J_t) takes its values in a compact subset \mathcal{H}_0 of the set of matrices (G, J) such that JJ^* and $GS(J)G^*$ are definite positive; we have also supposed that \bar{P}_0 is in some compact subset \mathcal{H}_1 of M_d^+ . On the other hand, $\Lambda(G, J, p)$ is uniformly bounded for $(G, J) \in \mathcal{H}_0$, $p \in \mathcal{H}_1$, and $\inf_{\mathcal{H}_0} \Lambda(G, J, p)$ converges to $+\infty$ as p tends to $\partial M_d^+ \cup \{\infty\}$. We deduce that there exists a compact subset \mathcal{H} of M_d^+ such that

$$(3.21) \quad \inf_{p \notin \mathcal{H}} \inf_{(G, J) \in \mathcal{H}_0} \Lambda(G, J, p) > \sup_{p \in \mathcal{H}_1} \sup_{(G, J) \in \mathcal{H}_0} \Lambda(G, J, p).$$

Consider the subdivision $s_i = i\varepsilon$ and let τ be the infimum of times $t \geq 0$ such that $\bar{P}_t \notin \mathcal{H}$ or $|\bar{P}_t - \bar{P}_{s_i}| \geq \varepsilon^{1/4}$ for $s_i \leq t < s_{i+1}$. In Lemma 6, we prove the estimates (1.35) on the random observable time interval $[0, \tau]$; this implies that, for T fixed, these estimates hold on the time interval $[0, T]$ except on the event $\{T < \tau\}$, so it is then sufficient to prove that the probability of this event is $O(\varepsilon^\infty)$.

LEMMA 6. *The estimates about $H_t - \bar{H}_t$ and $P_t - \bar{P}_t$ stated in Theorem 3 hold on $\{t \leq \tau\}$.*

Proof. We deduce from part (3) of Lemma 1 that $-\bar{P}_t$ is exponentially stable in the timescale ε up to time τ , so the estimate about $H_t - \bar{H}_t$ follows from Theorem 1. Moreover, from Lemma 2, we have

$$(3.22) \quad \bar{P}_{t_{i+1}} - \bar{P}_{t_i} = -\frac{K}{2} \varepsilon^{1/4} \nabla \Lambda(G_{t_i}, J_{t_i}, \bar{P}_{t_i}) + O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}).$$

The function $\Lambda(G, J, p)$ is not defined for any matrix p , but we can extend it into a function $\Lambda_0(G, J, p)$ defined everywhere, which coincides with Λ for p in \mathcal{H} , tending to $+\infty$ at infinity, and such that $\nabla \Lambda_0$ is zero only at $p = P(G, J)$. Then $Z_t = \bar{P}_t - P_t$ satisfies the recursive condition (3.1) of Lemma 5 with $\beta_t = (G_t, J_t)$ and

$$(3.23) \quad b(G, J, z) = -\frac{K}{2} \nabla \Lambda_0(G, J, z + P(G, J)).$$

For any (G, J) , the flow defined by (3.2) is the flow of a gradient dynamical system with a globally attractive equilibrium, so, to apply Lemma 5, we must only check that the eigenvalues of $b'(G, J, 0)$ are not on the imaginary axis. Since $b'(G, J, 0)$ is symmetric, it is sufficient to prove that it is invertible. Thus, let us study the function $\Lambda(G, J, \bar{P})$ in the neighbourhood of $P(G, J)$; since $\nabla \bar{Q}(G, J, \bar{P})$ is zero for $\bar{P} = P(G, J)$, we have

$$(3.24) \quad \bar{Q}(G, J, \bar{P}) = Q(G, J) + O(|\bar{P} - P|^2).$$

From (1.17) and since $R(G, J, P) = 0$, we have

$$(3.25) \quad \begin{aligned} R(G, J, \bar{P}) &= Q(G, J) + GJ^* - \bar{P}JJ^* + O(|\bar{P} - P|^2) \\ &= (P - \bar{P})JJ^* + O(|\bar{P} - P|^2). \end{aligned}$$

From (1.18), we deduce that

$$(3.26) \quad \nabla \Lambda(G, J, \bar{P}) = -P^\dagger(P - \bar{P})JJ^* + O(|\bar{P} - P|^2),$$

so $b'(G, J, 0)$ is given by

$$(3.27) \quad b'(G, J, 0): z \mapsto -\frac{K}{2} P^\dagger zJJ^*.$$

In particular, it is invertible; so we can apply Lemma 5, and the estimate on $P_t - \bar{P}_t$ follows. \square

Proof of Theorem 3. As was stated previously, it is now sufficient to prove that, for any fixed T , the probability of $\{T > \tau\}$ is $O(\varepsilon^\infty)$; to this end, we must only prove that, for any i ,

$$(3.28) \quad \mathbb{P}[s_i \leq \tau < s_{i+1}] = O(\varepsilon^\infty)$$

uniformly in i . The estimation of $\mathbb{P}[T > \tau]$ then follows by adding this estimate over the $O(1/\varepsilon)$ possible values of i . On the event $\{s_i \leq \tau < s_{i+1}\}$, the variable $\bar{P}_\tau - \bar{P}_{s_i}$ is of order $\sqrt{\varepsilon}$, so

$$(3.29) \quad \begin{aligned} & \mathbb{P}[|\bar{P}_\tau - \bar{P}_{s_i}| \geq \varepsilon^{1/4}, s_i \leq \tau < s_{i+1}] \\ & \leq \varepsilon^{-k/4} \mathbb{E}[|\bar{P}_\tau - \bar{P}_{s_i}|^k, s_i \leq \tau < s_{i+1}] = O(\varepsilon^{k/4}) \end{aligned}$$

for any k and is therefore $O(\varepsilon^\infty)$. On the other hand, let p_t be the solution of

$$(3.30) \quad \dot{p}_t = -\frac{K}{2\sqrt{\varepsilon}} \nabla \Lambda(G_0, J_0, p_t), \quad p_0 = \bar{P}_0.$$

Since p_t is solution of a gradient equation, it follows that $\Lambda(G_0, J_0, p_t)$ is nonincreasing, so

$$(3.31) \quad \begin{aligned} \inf_{(G,J) \in \mathcal{H}_0} \Lambda(G, J, p_t) & \leq \Lambda(G_0, J_0, p_t) \leq \Lambda(G_0, J_0, \bar{P}_0) \\ & \leq \sup_{p \in \mathcal{H}_1} \sup_{(G,J) \in \mathcal{H}_0} \Lambda(G, J, p). \end{aligned}$$

Since P_t is the point that minimizes $\Lambda(G_t, J_t, \cdot)$, we also have

$$(3.32) \quad \begin{aligned} \inf_{(G,J) \in \mathcal{H}_0} \Lambda(G, J, P_t) & \leq \Lambda(G_t, J_t, P_t) \leq \Lambda(G_t, J_t, \bar{P}_0) \\ & \leq \sup_{p \in \mathcal{H}_1} \sup_{(G,J) \in \mathcal{H}_0} \Lambda(G, J, p). \end{aligned}$$

Thus it follows from condition (3.21) that, on $\{t \leq \tau\}$, P_t and p_t take their values in \mathcal{H} , and that the distance between these two processes and $\partial\mathcal{H}$ is bounded below by a positive constant c_0 . From the application of Lemma 5 used in Lemma 6, we have

$$(3.33) \quad \bar{P}_t = P_t - P_0 + p_t + O(\varepsilon^{1/4})$$

on $\{t \leq \tau\}$. It follows from (3.30) that p_t converges exponentially fast to P_0 ; thus, if $s_i \geq \varepsilon^{1/4}$, then $P_{s_i} - \bar{P}_{s_i}$ is of order $\varepsilon^{1/4}$ on $\{s_i \leq \tau\}$; so $P_\tau - \bar{P}_\tau$ is of order $\varepsilon^{1/4}$ on $\{s_i \leq \tau < s_{i+1}\}$, and therefore

$$(3.34) \quad \begin{aligned} & \mathbb{P}[\bar{P}_\tau \in \partial\mathcal{H}, s_i \leq \tau < s_{i+1}] \\ & \leq \mathbb{P}[|\bar{P}_\tau - P_\tau| \geq c_0, s_i \leq \tau < s_{i+1}] = O(\varepsilon^\infty). \end{aligned}$$

Similarly, if $s_i < \varepsilon^{1/4}$, then $P_{s_i} - P_0$ is $O(\varepsilon^{1/8})$; so $p_{s_i} - \bar{P}_{s_i}$ is of order $\varepsilon^{1/8}$ on $\{s_i \leq \tau\}$, and therefore we obtain the same estimate. The wanted estimate (3.28) is then obtained by adding the two probabilities (3.29) and (3.34). \square

3.3. Locally stable filters and Theorem 4. For the proof of Theorem 4, let \mathcal{H} be a compact subset of M_d^+ such that P_t takes its values in a compact subset of the interior of \mathcal{H} and let τ be (as before) the infimum of times t such that $\bar{P}_t \notin \mathcal{H}$ or $|\bar{P}_t - \bar{P}_{s_i}| \geq \varepsilon^{1/4}$ for $s_i \leq t < s_{i+1}$. Let τ_0 be the infimum of times $t \geq 0$ such that $|P_t - \bar{P}_t| \geq \varepsilon^{1/8}$ or $|\bar{P}_t - \bar{P}_{s_i}| \geq \varepsilon^{1/4}$ for $s_i \leq t < s_{i+1}$. Note that τ_0 is not observable and that $\tau_0 \leq \tau$ for ε small enough.

LEMMA 7. *Under the conditions of Theorem 4, $H_t - \bar{H}_t$ is $O(\sqrt{\varepsilon})$ on $\{t \leq \tau\}$ and $P_t - \bar{P}_t$ is $O(\varepsilon^{1/4})$ on $\{t \leq \tau_0\}$.*

Proof. From part (3) of Lemma 1, $-\bar{P}_t$ is exponentially stable up to time τ , so we can apply Theorem 1 and obtain the estimate on $H_t - \bar{H}_t$. By applying (3.26), it appears that

$$(3.35) \quad K_t[\nabla \Lambda(G_t, J_t, \bar{P}_t)] = \bar{\rho}_t[\bar{P}_t - P_t]$$

for some process $\bar{\rho}_t$ that is a smooth function of K_t, G_t, J_t, \bar{P}_t such that

$$(3.36) \quad |\bar{\rho}_t - \rho(K_t, \bar{P}_t^\dagger, J_t J_t^*)| \leq C|\bar{P}_t - P_t|.$$

The Lipschitz dependence of $\bar{\rho}_t$ implies that the oscillations of $\bar{\rho}_t$ are $O(\varepsilon^{1/4})$ on $[t_i, t_{i+1}]$, and we deduce from part (2) of Lemma 1 and (3.36) that $-\bar{\rho}_t$ is exponentially stable in the timescale $\sqrt{\varepsilon}$ up to time τ_0 . On the other hand, from Lemma 2 and (3.35),

$$(3.37) \quad \bar{P}_{t_{i+1}} = \bar{P}_{t_i} - \frac{\varepsilon^{1/4}}{2} \bar{\rho}_{t_i} [\bar{P}_{t_i} - P_{t_i}] + O_t(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}).$$

Therefore, we can apply Lemma 4 to $Z_t = P_t - \bar{P}_t$ and deduce that this process is of order $\varepsilon^{1/4}$ on $\{t \leq \tau_0\}$. \square

Proof of Theorem 4. Let us consider the event $\{s_i \leq \tau_0 < s_{i+1}\}$; we deduce from Lemma 7 that, on this event, $\bar{P}_{\tau_0} - P_{\tau_0}$ and $\bar{P}_{\tau_0} - \bar{P}_{s_i}$ are, respectively, of order $\varepsilon^{1/4}$ and $\sqrt{\varepsilon}$. Thus, by proceeding as in Theorem 3, the probability of this event is $O(\varepsilon^\infty)$. By summing over the $O(1/\varepsilon)$ possible values of i , we deduce that $\mathbb{P}[\tau_0 < T]$ is $O(\varepsilon^\infty)$. On the other hand, note that $P_t - \bar{P}_t$ is $O(1)$ on $\{t \leq \tau\}$, is $O(\varepsilon^{1/4})$ on $\{t \leq \tau_0\}$, and that the probability of $\{\tau_0 < t \leq \tau\}$ is $O(\varepsilon^\infty)$; we easily deduce that $P_t - \bar{P}_t$ is $O(\varepsilon^{1/4})$ on $\{t \leq \tau\}$. Moreover, since $\tau_0 \leq \tau$, the probability of $\{\tau < T\}$ is $O(\varepsilon^\infty)$, so we can conclude as in Theorem 3. \square

3.4. Linearization and Theorem 5. For the estimation of \hat{H}_t , the idea is to choose some time before t and to compare our system with the linear system obtained by freezing the parameters at that time. This comparison is based on a change of probability chosen, so that, under the new probability, Y_t becomes the observation process of a linear system.

Proof of Theorem 5. Put $t'_i = i\varepsilon^{8/9}$, let us fix some time t , and let i be the index such that $t'_{i+1} \leq t < t'_{i+2}$. Let τ be the infimum of times $s \geq t'_i$ such that $|H_s - \bar{H}_s| \geq 1$, $|P_s - \bar{P}_s| \geq 1$, $|G_s - G_{t'_i}| \geq \varepsilon^{1/9}$, or $|J_s - J_{t'_i}| \geq \varepsilon^{1/9}$. Then, for $t'_i \leq s \leq t$, $H_s - \bar{H}_s$ and $P_s - \bar{P}_s$ are, respectively, $O(\sqrt{\varepsilon})$ and $O(\varepsilon^{1/4})$ on $\{t \leq \tau\}$, and the probability of this event is $O(\varepsilon^\infty)$. Define

$$(3.38) \quad \bar{W}_s = \int_0^s (G_u S_u G_u^*)^{-1/2} G_u S_u dW_u, \quad B_s = \int_0^s J_u dW_u.$$

They are two independent standard Wiener processes with respective dimensions n and d , and the system for (H, Y) can be written as

$$(3.39) \quad dH_s = F_s ds + (G_s S_s G_s^*)^{1/2} d\bar{W}_s + G_s J_s^* dB_s,$$

$$(3.40) \quad dY_s = H_s ds + \varepsilon dB_s.$$

Now consider the processes $\tilde{H}_s, \tilde{B}_s, \tilde{W}_s$, which coincide with H_s, B_s, \bar{W}_s for $s \leq t'_i$ and which are solutions of

$$(3.41) \quad d\tilde{H}_s = (G_{t'_i} S_{t'_i} G_{t'_i}^*)^{1/2} d\tilde{W}_s + G_{t'_i} J_{t'_i}^* d\tilde{B}_s,$$

$$(3.42) \quad d\tilde{B}_s = dB_s + \frac{1}{\varepsilon} (H_s - \tilde{H}_s) ds,$$

$$(3.43) \quad d\tilde{W}_s = d\bar{W}_s + \frac{1}{\varepsilon} (G_s S_s G_s^*)^{-1/2} G_s J_s^* (\tilde{H}_s - H_s) ds$$

for $s \geq t'_i$; this is a linear equation in \tilde{H} , so, in particular, it has a unique solution. Note that, from (3.40) and (3.42), the observation is given by

$$(3.44) \quad dY_s = \tilde{H}_s ds + \varepsilon d\tilde{B}_s.$$

Consider also the local martingale given by $L_s = 1$ for $s \leq t'_i$ and

$$(3.45) \quad dL_s = \frac{L_s}{\varepsilon} (\tilde{H}_s - H_s)^* dB_s - \frac{L_s}{\varepsilon} (\tilde{H}_s - H_s)^* J_s G_s^* (G_s S_s G_s^*)^{-1/2} d\bar{W}_s$$

for $s \geq t'_i$. For ε fixed and s bounded, we deduce from (1.40) and

$$(3.46) \quad d(H_s - \tilde{H}_s) = F_s ds + \frac{1}{\varepsilon} ((G_{t'_i} S_{t'_i} G_{t'_i}^*)^{1/2} (G_s S_s G_s^*)^{-1/2} G_s J_s^* - G_{t'_i} J_{t'_i}^*) (H_s - \tilde{H}_s) ds \\ + ((G_s S_s G_s^*)^{1/2} - (G_{t'_i} S_{t'_i} G_{t'_i}^*)^{1/2}) d\bar{W}_s + (G_s J_s^* - G_{t'_i} J_{t'_i}^*) dB_s$$

that an exponential moment of $|H_s - \tilde{H}_s|^2$ is bounded, so it follows from classical results that L_s is actually a martingale. Thus we can consider the probability $\tilde{\mathbb{P}} = L_{t_i} \mathbb{P}$ on \mathcal{F}_{t_i} . Under this probability, \tilde{W}_s and \tilde{B}_s are independent Wiener processes. We now have two probabilities; note, however, that the notation $O(\varepsilon^k)$ is a relation of domination in the spaces $L^k(\mathbb{P})$, but not necessarily in $L^k(\tilde{\mathbb{P}})$. Consider the process

$$(3.47) \quad h_s = \tilde{\mathbb{E}}[\tilde{H}_s | \mathcal{F}_{t'_i} \vee \mathcal{Y}_s].$$

Under $\tilde{\mathbb{P}}$, after time t'_i , the filtering problem (3.41)–(3.44) is linear with coefficients $(G_{t'_i}, J_{t'_i})$, so the optimal filter is given by a Kalman filter

$$(3.48) \quad dh_s = \frac{p_s}{\varepsilon} (dY_s - h_s ds), \quad h_{t'_i} = H_{t'_i},$$

where the gain p_s is solution of some differential Riccati equation. The solution of this equation converges exponentially fast to the solution of the algebraic Riccati equation, which is $P_{t'_i}$. Since $P_s - \bar{P}_s$ is $O(\varepsilon^{1/4})$ up to time τ , we deduce that, for $t'_i \leq s \leq t$,

$$(3.49) \quad p_s = P_{t'_i} + O(e^{-c(s-t'_i)/\varepsilon}) = P_s + O(\varepsilon^{4/9} + e^{-c(s-t'_i)/\varepsilon}) \\ = \bar{P}_s + O(\varepsilon^{1/4} + e^{-c(s-t'_i)/\varepsilon})$$

on $\{t \leq \tau\}$. Moreover, the error covariance matrix for the linear filter (3.48) is εp_s ; so

$$(3.50) \quad \tilde{\mathbb{E}}[|\tilde{H}_s - h_s|^k | \mathcal{F}_{t'_i} \vee \mathcal{Y}_s]^{1/k} = \tilde{\mathbb{E}}[|\tilde{H}_s - h_s|^k]^{1/k} = O(\sqrt{\varepsilon})$$

for any k . By comparing the equations (3.48) of h_s and (1.8) of \tilde{H}_s , and by applying (3.49) and part (4) of Lemma 1, since $t - t'_i \geq \varepsilon^{8/9} \gg \varepsilon$, it appears that $h_t - \tilde{H}_t$ is $O(\varepsilon^{3/4})$ on $\{t \leq \tau\}$. On the other hand, note from (3.46) that $H_s - \tilde{H}_s$ is $O(\varepsilon^{8/9})$ on $\{t \leq \tau\}$, so $\ln L_{t_i}$ is $O(\varepsilon^{1/3})$. Thus the event

$$(3.51) \quad \Omega_0 = \{|L_t - 1| \leq \varepsilon^{2/7}\} \cap \{t \leq \tau\}$$

is of probability $1 - O(\varepsilon^\infty)$ under \mathbb{P} . We deduce that

$$(3.52) \quad \tilde{\mathbb{P}}[\Omega_0] = \mathbb{E}[L_t 1_{\Omega_0}] \geq (1 - \varepsilon^{2/7}) \mathbb{P}[\Omega_0] \geq 1 - O(\varepsilon^{2/7}),$$

so, for any event A ,

$$(3.53) \quad \tilde{\mathbb{P}}[A] = \tilde{\mathbb{P}}[A \cap \Omega_0] + \tilde{\mathbb{P}}[A \cap \Omega_0^c] = \mathbb{E}[L_t 1_{A \cap \Omega_0}] + O(\varepsilon^{2/7}) \\ = \mathbb{P}[A] + O(\varepsilon^{2/7})$$

uniformly in A . This implies that a process is of order ε^k in \mathbb{P} -probability in the sense of (1.3) if and only if it has the same property in $\tilde{\mathbb{P}}$ -probability, and we will use the notation $O_p(\varepsilon^k)$ for this relation of domination. We have already checked that $\tilde{H}_t - h_t$ and $H_t - \tilde{H}_t$ are, respectively, $O(\varepsilon^{3/4})$ and $O(\varepsilon^{8/9})$ on $\{t \leq \tau\}$, so we have

$$(3.54) \quad \hat{H}_t - \bar{H}_t = \mathbb{E}[H_t - \bar{H}_t | \mathcal{Y}_t] = \mathbb{E}[(\tilde{H}_t - h_t) 1_{\Omega_0} | \mathcal{Y}_t] + O_p(\varepsilon^{3/4}).$$

Let \mathbb{P}_0 be the probability \mathbb{P} conditioned on Ω_0 ; then

$$\begin{aligned} \mathbb{E}[(\tilde{H}_t - h_t)1_{\Omega_0}|\mathcal{Y}_t] &= \mathbb{E}_0[\tilde{H}_t - h_t|\mathcal{Y}_t]\mathbb{P}[\Omega_0|\mathcal{Y}_t] \\ (3.55) \qquad \qquad \qquad &= \mathbb{E}_0[\tilde{H}_t - h_t|\mathcal{Y}_t] + O_p(\varepsilon^\infty), \end{aligned}$$

so

$$\begin{aligned} \hat{H}_t - \bar{H}_t &= \mathbb{E}_0[\tilde{H}_t - h_t|\mathcal{Y}_t] + O_p(\varepsilon^{3/4}) \\ (3.56) \qquad \qquad \qquad &= \frac{\tilde{\mathbb{E}}[L_t^{-1}1_{\Omega_0}(\tilde{H}_t - h_t)|\mathcal{Y}_t]}{\tilde{E}[L_t^{-1}1_{\Omega_0}|\mathcal{Y}_t]} + O_p(\varepsilon^{3/4}). \end{aligned}$$

Note that

$$(3.57) \qquad \qquad \qquad |L_t^{-1}1_{\Omega_0} - 1| \leq \varepsilon^{2/7} + 1_{\Omega_0^c},$$

so, from (3.52),

$$(3.58) \qquad \qquad \qquad \tilde{\mathbb{E}}[|L_t^{-1}1_{\Omega_0} - 1|^{8/7}] \leq C\varepsilon^{2/7}.$$

From (3.50), we also verify that

$$(3.59) \qquad \qquad \qquad \tilde{\mathbb{E}}[|\tilde{H}_t - h_t|^8|\mathcal{Y}_t]^{1/8} = \tilde{\mathbb{E}}[|\tilde{H}_t - h_t|^8]^{1/8} = O(\sqrt{\varepsilon}),$$

so, from Hölder's inequality,

$$\begin{aligned} \tilde{\mathbb{E}}[\tilde{\mathbb{E}}[L_t^{-1}1_{\Omega_0}(\tilde{H}_t - h_t)|\mathcal{Y}_t]] &\leq \tilde{\mathbb{E}}[(L_t^{-1}1_{\Omega_0} - 1)(\tilde{H}_t - h_t)| \\ (3.60) \qquad \qquad \qquad &= O(\varepsilon^{1/4})O(\sqrt{\varepsilon}) = O(\varepsilon^{3/4}). \end{aligned}$$

Thus the numerator of (3.56) is $O_p(\varepsilon^{3/4})$; the denominator is $1 + O_p(\varepsilon^{2/7})$, so the quotient is $O_p(\varepsilon^{3/4})$. \square

4. Proof of Theorem 6. Consider the subdivision $s'_i = i\sqrt{\varepsilon}$ and let τ be the infimum of times $t \geq 0$ such that $|\bar{X}_t - \bar{X}_{s'_i}| \geq \varepsilon^{1/8}$ for $s'_i \leq t < s'_{i+1}$ or $|X_t - \bar{X}_t| \geq \varepsilon^{3/16}$.

LEMMA 8. *Under the conditions of Theorem 6, the estimates on $X_t - \bar{X}_t$ and $h(X_t) - h(\bar{X}_t)$ hold on $\{t \leq \tau\}$.*

Proof. First, we prove from part (3) of Lemma 1 that $-p(\bar{X}_t)$ is exponentially stable up to time τ in the timescale ε , so, by applying part (4) of Lemma 1 to $h(X_t) - h(\bar{X}_t)$, it appears that this process is $O(\sqrt{\varepsilon})$ on $\{t \leq \tau\}$. Define

$$(4.1) \qquad \qquad \qquad \gamma = I - \phi p^{-1}h'.$$

For each x , $\gamma(x)$ is the projection on $\ker h'(x)$ in the direction of the image of $\phi(x)$. Observe that

$$\begin{aligned} X_t - \bar{X}_t &= \gamma(\bar{X}_t)(X_t - \bar{X}_t) + \phi p^{-1}(\bar{X}_t)(h(X_t) - h(\bar{X}_t)) \\ (4.2) \qquad \qquad \qquad &- \phi p^{-1}(\bar{X}_t)(h(X_t) - h(\bar{X}_t) - h'(\bar{X}_t)(X_t - \bar{X}_t)), \end{aligned}$$

so if we define

$$(4.3) \qquad \qquad \qquad Z_t = \gamma(\bar{X}_t)(X_t - \bar{X}_t),$$

since $|X_t - \bar{X}_t|^2$ and $h(X_t) - h(\bar{X}_t)$ are, respectively, $O(\varepsilon^{3/8})$ and $O(\sqrt{\varepsilon})$, we obtain

$$(4.4) \qquad \qquad \qquad X_t - \bar{X}_t = Z_t + O(\varepsilon^{3/8})$$

on $\{t \leq \tau\}$. Therefore, we must prove that Z_t is $O(\varepsilon^{1/4})$ on $\{t \leq \tau\}$. From the definition of γ and the properties of ϕ and ψ , we have $\gamma\phi = 0$ and $\gamma\psi = \psi$, so, by applying Ito's formula to (4.3),

$$(4.5) \quad \begin{aligned} dZ_t &= \gamma(\bar{X}_t) dX_t + d\gamma(\bar{X}_t)(X_t - \bar{X}_t) + d\langle \gamma(\bar{X}), X - \bar{X} \rangle_t \\ &\quad - \frac{1}{\varepsilon^{3/2}} \psi(\bar{X}_t) [\Gamma_t^*(dY_t - h(\bar{X}_t)dt)]. \end{aligned}$$

Since

$$(4.6) \quad d\bar{X}_t = O(\varepsilon^{-1/2})dt + O(1)dW_t,$$

the oscillations of \bar{X}_t on $[t_i, t_{i+1}]$ are $O(\varepsilon^{1/4})$, so Lemma 2 shows that

$$(4.7) \quad \varepsilon^{-7/4} \int_{t_i}^{t_{i+1}} \Gamma_t^*(dY_t - h(\bar{X}_t)dt) = \frac{1}{2} p^+(\bar{X}_{t_i}) r(X_{t_i}, \bar{X}_{t_i}) + O_i(\varepsilon^{1/8}) + O(\varepsilon^{1/4}),$$

where, from (3.25),

$$(4.8) \quad r(x, \bar{x}) = R(g(x), j(x), p(\bar{x})) = (p'(\bar{x})(x - \bar{x}))jj^*(\bar{x}) + O(|x - \bar{x}|^2).$$

Thus, from (4.4),

$$(4.9) \quad r(X_t, \bar{X}_t) = (p'(\bar{X}_t)Z_t)jj^*(\bar{X}_t) + O(\varepsilon^{3/8}),$$

and since the oscillations of $\psi(\bar{X}_t)$ are $O(\varepsilon^{1/4})$, we deduce from (4.7) and the definition (1.42) of ρ that

$$(4.10) \quad \varepsilon^{-7/4} \int_{t_i}^{t_{i+1}} \psi(\bar{X}_t) [\Gamma_t^*(dY_t - h(\bar{X}_t)dt)] = \frac{1}{2} \rho(\bar{X}_{t_i})Z_{t_i} + O_i(\varepsilon^{1/8}) + O(\varepsilon^{1/4}).$$

On the other hand, since the oscillations of $X_t - \bar{X}_t$ are $O(\varepsilon^{1/4})$,

$$(4.11) \quad \int_{t_i}^{t_{i+1}} d\gamma(\bar{X}_u)(X_u - \bar{X}_u) = (\gamma(\bar{X}_{t_{i+1}}) - \gamma(\bar{X}_{t_i}))(X_{t_i} - \bar{X}_{t_i}) + O(\sqrt{\varepsilon}),$$

so, from (4.5) and (4.10),

$$(4.12) \quad \begin{aligned} Z_{t_{i+1}} - Z_{t_i} &= -\frac{\varepsilon^{1/4}}{2} \rho(\bar{X}_{t_i})Z_{t_i} + (\gamma(\bar{X}_{t_{i+1}}) \\ &\quad - \gamma(\bar{X}_{t_i}))(X_{t_i} - \bar{X}_{t_i}) + O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}). \end{aligned}$$

The variables Z_{t_i} , $\bar{X}_{t_{i+1}} - \bar{X}_{t_i}$, and $X_{t_i} - \bar{X}_{t_i}$ are, respectively, of order $\varepsilon^{3/16}$, $\varepsilon^{1/4}$ and $\varepsilon^{3/16}$ on $\{t \leq \tau\}$, so $Z_{t_{i+1}} - Z_{t_i}$ is of order $\varepsilon^{3/8}$, and, from (4.4), $\bar{X}_{t_{i+1}} - \bar{X}_{t_i}$ has the same order. This implies that the second term in the estimation (4.12) is actually $O(\varepsilon^{9/16})$. Thus

$$(4.13) \quad Z_{t_{i+1}} - Z_{t_i} = -\frac{\varepsilon^{1/4}}{2} \rho(\bar{X}_{t_i})Z_{t_i} + O_i(\varepsilon^{3/8}) + O(\sqrt{\varepsilon}).$$

Define

$$(4.14) \quad \bar{\rho} = \rho\gamma + \phi p^{-1}h'.$$

Then the linear map $\bar{\rho}(x)$ is equal to $\rho(x)$ on $\ker h'(x)$ and to I on the image of $\phi(x)$; thus we can replace ρ by $\bar{\rho}$ in the estimation (4.13), and the set of eigenvalues of $\bar{\rho}(x)$ is equal to the set of eigenvalues of $\rho(x)$ in $\ker h'(x)$, plus the value 1. In particular, the real parts of these eigenvalues are bounded below by a positive constant; moreover,

the oscillations of \bar{X}_t are controlled, so we can deduce from part (3) of Lemma 1 that $-\bar{\rho}(\bar{X}_t)$ is exponentially stable in the timescale $\sqrt{\varepsilon}$. Then it follows from Lemma 4 that Z_t is $O(\varepsilon^{1/4})$ on $\{t \leq \tau\}$, so $X_t - \bar{X}_t$ satisfies the same property. \square

Proof of Theorem 6. On $\{t \leq \tau\}$, $X_t - \bar{X}_t$ and

$$(4.15) \quad \bar{X}_t - \bar{X}_{s'_t} = (\bar{X}_t - X_t) - (\bar{X}_{s'_t} - X_{s'_t}) + (X_t - X_{s'_t})$$

are $O(\varepsilon^{1/4})$, so we verify as in Theorem 3 or Theorem 4 that the probability of $\{T > \tau\}$ is $O(\varepsilon^\infty)$. The estimate on \hat{H}_t follows from Theorem 5. \square

REFERENCES

- [1] A. BENSOUSSAN, *On some approximation techniques in nonlinear filtering theory*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. in Math. and Appl. 10, Springer, Berlin, New York, 1988.
- [2] A. BENVENISTE, M. MÉTIVIER, AND P. PRIOURET, *Algorithmes stochastiques et approximations stochastiques*, Masson, Paris, 1987.
- [3] W. H. FLEMING AND E. PARDOUX, *Piecewise monotone filtering with small observation noise*, SIAM J. Control Optim., 27 (1989), pp. 1156–1181.
- [4] YU. M. KABANOV, S. M. PERGAMENSHCHIKOV, AND J. M. STOYANOV, *Asymptotic expansions for singularly perturbed stochastic differential equations*, in Probability and Statistics (papers in honor of Yu. Prohorov), VNU Science Press.
- [5] R. KATZUR, B. Z. BOBROVSKY, AND Z. SCHUSS, *Asymptotic analysis of the optimal filtering problem for one-dimensional diffusions measured in a low noise channel, part II*, SIAM J. Appl. Math., 44 (1984), pp. 1176–1191.
- [6] A. J. KRENER, *The asymptotic approximation of nonlinear filters by linear filters*, in Theory and Applications of Nonlinear Control Systems, Stockholm, 1985, Elsevier, North-Holland, Amsterdam, 1986.
- [7] H. J. KUSHNER, *A projected stochastic approximation method for adaptive filters and identifiers*, IEEE Trans. Automat. Control, 25 (1980), pp. 836–838.
- [8] R. LIPSTER, AND J. STOYANOV, *Stochastic version of the averaging principle for diffusion type processes*, Stochastics Stochastics Rep., 32 (1990), pp. 145–163.
- [9] L. LJUNG, *Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 36–50.
- [10] E. MAYER-WOLF, M. ZAKAI, AND O. ZEITOUNI, *On the memory length of the optimal nonlinear filter*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. in Math. and Appl. 10, Springer, Berlin, New York, 1988.
- [11] P. MILHEIRO DE OLIVEIRA, *Etudes asymptotiques en filtrage non linéaire avec petit bruit d'observation*, Thèse de Doctorat, Univ. de Provence, France, 1990.
- [12] J. PICARD, *Nonlinear filtering of one-dimensional diffusions in the case of a high signal-to-noise ratio*, SIAM J. Appl. Math., 46 (1986), pp. 1098–1125.
- [13] ———, *Nonlinear filtering and smoothing with high signal-to-noise ratio*, in Stochastic Processes in Physics and Engineering, Bielefeld, 1986, Reidel, Boston, 1988.
- [14] ———, *Efficiency of the extended Kalman filter for nonlinear systems with small noise*, SIAM J. Appl. Math., 51 (1991), pp. 843–885.
- [15] ———, *A nonlinear filter with two time scales*, in Proc. Workshop on Applied Stochastic Analysis, Rutgers Univ., New Brunswick, NJ, 1991, Lecture Notes on Control Inform. Sci., 177, Springer, Berlin, New York, 1992.
- [16] J. H. VAN SCHUPPEN, *Convergence results for continuous-time adaptive stochastic filtering algorithms*, J. Math. Anal. Appl., 96 (1983), pp. 209–225.
- [17] Y. TAKEUCHI AND H. AKASHI, *Least-squares state estimation of systems with state-dependent observation noise*, Automatica, 21 (1985), pp. 303–313.
- [18] I. YAESH, B. Z. BOBROVSKY, AND A. SCHUSS, *Asymptotic analysis of the optimal filtering problem for two dimensional diffusions measured in a low noise channel*, SIAM J. Appl. Math., 50 (1990), pp. 1134–1155.

CONVEX DUALITY AND NONLINEAR OPTIMAL CONTROL*

RICHARD VINTER†

This paper is dedicated to Wendell Fleming on the occasion of his 65th birthday.

Abstract. Problems in nonlinear optimal control can be reformulated as convex optimization problems over a vector space of linear functionals. In this way, methods of convex analysis can be brought to bear on the task of characterizing solutions to such problems. The result is a necessary and sufficient condition of optimality that generalizes well-known sufficient conditions, referred to as verification theorems, in dynamic programming; as a byproduct, we obtain a representation of the minimum cost in terms of the upper envelope of subsolutions to the Hamilton–Jacobi equation. It is a striking illustration of the wide range of problems to which convex analysis, and, in particular, convex duality, is applicable. The approach, applied to parametric problems in the calculus of variations, was pioneered by L. C. Young [*Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, PA, 1969]. As recent work has shown, however, it is equally fruitful when applied in optimal control. This paper, which is expository, offers a self-contained treatment of the application of methods of convex duality to general nonlinear problems in deterministic optimal control. At the same time, it provides extensions of previously published results in several directions. A simple proof is given of the main “convex closure” theorem relating generalized flows and relaxed arcs; this is based on mollification techniques recently developed by Fleming and Vermes [*SIAM J. Control Optim.*, 27 (1989), pp. 1136–1155] for constructing smooth subsolutions to the Hamilton–Jacobi equation.

Key words. optimal control, convex analysis, dynamic programming

AMS(MOS) subject classifications. 49C05, 49A55, 49A52, 90C25

Introduction. Convex analysis is widely acknowledged to have an important but specialized role in optimization. The class of convex optimization problems with which it is concerned is one for which global minimizers admit a simple and precise characterization. Excluding degenerate cases, we find that for such problems the usual multiplier rules are necessary and sufficient for global optimality, that multipliers can be interpreted as minimizers for dual maximization problems, and that the supremum of the cost for the dual problem provides a tight lower bound on the minimum cost for the original problem. These advantages are bought at a heavy price, however,—a theory for convex optimization problems casts aside at the outset the possible existence of local minima as well as other phenomena that are correctly associated with nonlinear systems.

In view of these associations, it comes as a surprise that convex analysis has a significant role in the study of *fully nonlinear* problems in the calculus of variations and optimal control. The marriage of special techniques and general problems is possible here because, even if an optimization problem with which we are initially presented involves a nonconvex cost and nonconvex constraints, there is scope for changing the linear structure on the underlying vector space (or embedding the domain in a new vector space, which is the same) to furnish a convex problem.

A way to harness methods of convex analysis for application to parametric problems in the calculus of variations is provided by L. C. Young’s theory of generalized flows, an expository version of which appears in [15]. Let us refer to our initial problem as the “strong” problem, and the convex problem with which we associate it as the “weak” problem. Setting up the weak problem involves embedding the class of arcs considered in the strong problem in a new vector space \mathcal{X} , some space of linear

* Received by the editors October 18, 1991; accepted for publication (in revised form) December 23, 1991.

† Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, Exhibition Road, London SW7 2BT, England.

functionals, and extending the cost function for the strong problem as a linear function over \mathcal{X} . The constraint set for the weak problem is a convex set, defined by a family of affine equality and inequality constraint functions, which contains the images of arcs for the original problem under the embedding. Conditions on a minimizer for the strong problem are then obtained by showing that it corresponds to a minimizer for the weak problem and by applying methods of convex analysis to this convex optimization problem.

Young's ideas were placed in a broader context by Ioffe [5], Levin and Milyutin [8], Klötzler [7], and Vinter [11] and extended to apply to nonlinear problems in optimal control. These papers typically establish existence, in some sense, of a subgradient of a certain convex function associated with the cost over some abstract vector space of boundaries. As shown by Vinter and Lewis [12], [13], connections between such results and other, more familiar ones in optimal control theory become apparent when methods of convex duality are applied to the weak problem. In this way, we obtain necessary and sufficient conditions of optimality, which improve on the familiar sufficient conditions passing under the name of "verification theorems" in the dynamic programming literature. A byproduct is a representation of the minimum cost in terms of the upper envelope of smooth subsolutions to the Hamilton–Jacobi equation. Recently, Fleming [3] and Fleming and Vermes [4] successfully adapted the approach to apply to a large class of optimal control problems, relating to both deterministic and stochastic systems, and they have simplified, by means of mollification techniques, proofs of the key intermediate steps relating the strong and the weak problems.

When we attempt to characterize minimizers in terms of the solutions to the Hamilton–Jacobi equation for general classes of optimal control problems, we encounter the difficulty that strict sense solutions (i.e., continuously differentiable ones) to the Hamilton–Jacobi equation may fail to exist. A number of remedies are available based on a variety of notions of generalized solution to the Hamilton–Jacobi equation (viscosity solutions [9], generalized solutions involving the Clarke generalized gradients [2], or lower Dini derivatives [14], among others). It is now apparent that the convex analysis approach provides another means of overcoming the difficulty—in place of strict sense solutions to the Hamilton–Jacobi equation, optimality conditions are given in terms of smooth subsolutions to the equation.

A distinguishing feature of the convex duality approach is that it may be applied to optimal control problems with general endpoint constraints, even in "degenerate" situations. Optimality conditions supplied in [2], which also treat endpoint constraints, are typical in requiring satisfaction of certain "calmness" or "normality" hypotheses. On the other hand, the viscosity solutions literature is primarily concerned with situations where the value function associated with perturbations of the initial time and state is defined on some suitably large domain, situations corresponding to no endpoint constraints, where nondegeneracy hypotheses of a "controllability" nature apply or where we place restrictions on the manner in which arcs may strike the target set. Such hypotheses play no part here.

Another feature is the mild nature of the hypotheses under which optimality conditions can be derived by means of convex duality. It is required that the cost be expressible in terms of lower semicontinuous functions and the multifunction, by means of which the dynamics are modelled, be upper semicontinuous. We are not limited, in particular, to treating problems for which the data is Lipschitz continuous in the state variable, as in [2], [14].

This paper is expository in nature; the arguments involved are direct, and the proofs are largely self-contained. The proof of the central "convex closure" theorem

is based on techniques of Fleming [3] and Fleming and Vermes [4] for constructing smooth subsolutions to the Hamilton–Jacobi equation, which take a particularly simple form for the deterministic problems considered here. At the same time, we provide extensions in several directions. As compared with [12], [13], a more general cost function is now permitted, requirements that the data is continuous are weakened, and the hypothesis implicit in [13] that all points in the target set be reachable is completely dispensed with. Changes, too, in the duality arguments arising in the analysis of the weak problem lead to improved optimality conditions, as indicated above, and, once again, simpler proofs. The differentiability and linear growth conditions on the data, implicit in the hypotheses of [4] (specialized to the deterministic case), are replaced by “semicontinuity” conditions and boundedness assumptions on the underlying domain.

1. Generalized flows. Our object here is to set up machinery for the application of methods of convex duality to nonlinear problems of optimal control. This involves abstracting the classical concept of an arc (“ordinary arcs”). Two levels of abstraction are involved. The first is to view it as a special case of a “relaxed arc.” This concept is a familiar one, and its role in existence theory is widely appreciated. The essential idea is to regard values of the velocity as linear functionals. The second level of abstraction is to interpret a relaxed arc itself as a linear functional, possessing certain arc-like properties, termed a generalized flow. It is this lesser-known concept that is the key to the reformulation of nonlinear optimal control problems as a convex optimization problems.

1.1. Relaxed arcs. A well-known phenomenon in the calculus of variations is where, if the cost integrand is not convex in the velocity variable, then minimizing sequences of arcs may have velocities that switch increasingly rapidly. To provide an existence theory covering such situations, we must attach meaning to “limits” of such sequences. For this purpose, Young introduced the concept of generalized curves, or “relaxed arcs” according to modern nomenclature. We give a brief review of the ideas involved, slanted toward the requirements of the following sections.

Take $[t_0, T]$ to be a fixed interval and $\Omega \subset \mathbb{R}^n$ a compact set containing the origin.

DEFINITION 1.1. An element (t_1, x, m) is a *relaxed arc* (with reference to the velocity set Ω) if $t_1 \in [t_0, T]$, $x(\cdot)$ is a Lipschitz continuous mapping from $[t_0, t_1]$ to \mathbb{R}^n , m (also written $t \rightarrow m_t$) maps $[t_0, t_1]$ into the space of regular Borel probability measures on Ω , and the following conditions are satisfied:

- (i) m is measurable in the sense that, for each $g \in C(\Omega)$, the scalar-valued function $t \rightarrow \int g(v)m_t(dv)$ is Lebesgue measurable;
- (ii) m is a “generalized velocity” of $x(t)$ in the sense that

$$\dot{x}(t) = \int vm_t(dv) \quad \text{a.e. } t \in [t_0, t_1].$$

The set of relaxed arcs can be regarded as an extension of the class of *ordinary arcs*, namely, elements (t_1, x) for which $t_1 \in [t_0, T]$, and x is a Lipschitz continuous \mathbb{R}^n -valued function on $[t_0, t_1]$ satisfying the velocity constraint “ $\dot{x}(t) \in \Omega$ almost everywhere.” To this end, we associate with an ordinary arc (t_1, x) the relaxed arc (t_1, x, m) for which

$$m_t := \delta_{\{\dot{x}(t)\}} \quad \text{a.e. } [t_0, t_1].$$

Here $\delta_{\{a\}}$ denotes the probability measure concentrated at the point $\{a\}$.

It is often convenient to regard the functions x and m , which constitute a relaxed arc (t_1, x, m) , as functions with domain all of $[t_0, T]$. For this purpose, we extend x

by constant extrapolation from its value at $t = t_1$; thus

$$x(t) := x(t_1) \quad \text{for } t \in (t_1, T],$$

and we extend m by setting

$$m_t := \delta_{\{0\}} \quad \text{for } t \in (t_1, T].$$

The following theorem is proved by methods of ([15, §67]). It refers to the set of integrands \mathcal{G}

$$\mathcal{G} := \{h : [t_0, T] \times \mathbb{R}^n \times \Omega \rightarrow \mathbb{R} : h(\cdot, y, v) \text{ is measurable, } h(t \cdot, \cdot) \text{ is continuous, and } t \rightarrow |h(t \cdot, \cdot)|_{C(B)} \text{ is integrable for any compact set } B \subset \mathbb{R}^{2n}\}.$$

THEOREM 1.2. *Let $\{(t_i, x_i, m^i)\}$ be a sequence of relaxed arcs such that the x_i 's are uniformly bounded in $C([t_0, T]; \mathbb{R}^n)$. Then there exists a subsequence (also written $\{(t_1^i, x_i, m^i)\}$) and a relaxed arc (t_1, x, m) such that*

$$t_1^i \rightarrow t_1,$$

$$x_i \rightarrow x \text{ uniformly on } [t_0, T],$$

$$\int_{t_0}^{t_1^i} \int h(t, x_i(t), v) dm_t^i(v) dt \rightarrow \int_{t_0}^{t_1} \int h(t, x(t), v) dm_t(v) dt \quad \text{as } i \rightarrow \infty,$$

for each $h \in \mathcal{G}$.

Given an integrand $h \in \mathcal{G}$, the integral functional

$$(1.1) \quad J(t_1, x) := \int_{t_0}^{t_1} h(t, x(t), \dot{x}(t)) dt$$

over ordinary arcs (t_1, x) extends to a functional J_r over relaxed arcs as follows:

$$(1.2) \quad J_r(t_1, x, m) := \int_{t_0}^{t_1} \int h(t, x(t), v) dm_t(v) dt$$

in a manner that is compatible with the way in which we identify the ordinary arcs with a subclass of the relaxed arcs.

Since no convexity conditions feature in the definition of the set \mathcal{G} , for any h drawn from \mathcal{G} we should not expect the problem

$$\text{“Minimize } J(t_1, x) \text{ over ordinary arcs } (t_1, x) \text{ for which } x \in R\text{”}$$

to have a solution. Here R is a closed bounded subset of $C([t_0, t_1]; \mathbb{R}^n)$, and J is the integral functional (1.1). (The constraint set R may incorporate a priori bounds that we are able to establish on candidates for a minimizer.) We can at least, however, ensure existence of solutions to the following problem in which the domain of J is extended to the relaxed arcs:

$$\text{“Minimize } J_r(t_1, x, m) \text{ over relaxed arcs } (t_1, x, m) \text{ for which } x \in R\text{.”}$$

Provided that the set of (t_1, x, m) 's satisfying $x \in R$ is nonempty, there exists a minimizing sequence $\{(t_1^i, x_i, m^i)\}$ of relaxed arcs for the minimization problem with extended domain. According to Theorem 1.2, we may extract a subsequence (also written $\{(t_1, x_i, m^i)\}$) and identify a relaxed arc (t_1, x, m) such that

$$\inf J_r = \lim_i J_r(t_1^i, x_i, m^i) = J_r(t_1, x, m)$$

and $x \in R$. Here $\inf J_r$ denotes the infimum cost for problem J_r . Evidently, (t, x, m) is a minimizer for the “relaxed” problem.

1.2. Generalized flows. The next step is to view relaxed arcs, themselves an extension of the notion of ordinary arc, as examples of “generalized flows.” Generalized flows are linear functions that capture certain significant properties of relaxed arcs.

Take a point $x_0 \in \mathbb{R}^n$ and an interval $[t_0, T]$. Let K and r be positive numbers that satisfy

$$(1.3) \quad |x_0| + r|T - t_0| \leq K.$$

We write

$$D = [t_0, T] \times (K\bar{B}) \quad \text{and} \quad \Omega = r\bar{B},$$

where \bar{B} denotes the closed unit ball. Consider now the following set of constraints:

$$(C) \quad \begin{aligned} t_1 &\in [t_0, T], \\ \dot{x}(t) &\in \Omega \quad \text{a.e. } t \in [t_0, t_1], \\ (t, x(t)) &\in D \quad \text{for all } t \in [t_0, t_1], \\ x(t_0) &= x_0. \end{aligned}$$

A relaxed arc (t_1, x, m) with velocity set Ω , which satisfies these constraints, will be called a *relaxed arc for system (C)*. (The second of these constraints is taken to indicate that Ω is the velocity set of the relaxed arc.) Evidently, the class of relaxed arcs for system (C) is nonempty.

A relaxed arc (t_1, x, m) for system (C) defines a linear functional (μ, γ) on $C(D \times \Omega) \times C(D)$ as follows:

$$(1.4) \quad \langle (\mu, \gamma), (\xi, \eta) \rangle = \int_{t_0}^{t_1} \int \xi(t, x(t), v) dm_t(v) dt + \eta(t_1, x(t_1))$$

for all $(\xi, \eta) \in C(D \times \Omega) \times C(D)$.

It is easy to see that (μ, γ) is, in fact, a bounded linear functional on $C(D \times \Omega) \times C(D)$ and hence defines an element in the vector space $C^*(D \times \Omega) \times C^*(D)$, which is isometrically isomorphic to the dual space $(C(D \times \Omega) \times C(D))^*$.

We denote by S the set of linear functionals that arise in this way. The symbol is chosen to indicate that they are associated with relaxed arcs (t_1, x) satisfying the constraints (C) in some “strong” sense as follows:

$$(1.5) \quad S := \{(\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D) : \text{there exists a relaxed arc for (C) satisfying (1.4)}\}.$$

The set S is nonempty; it contains the linear functional (μ, γ) associated with the ordinary arc $(T, x \equiv x_0)$, for example. S is obviously bounded with respect to the dual norm. It is also weak* compact. To see this, we first note that, by Theorem 1.2, S is sequentially weak* compact. S , however, is bounded in the dual norm, and $C(D \times \Omega) \times C(D)$ is separable. By Bishop’s theorem [16], the weak* topology on $C^* \times C^*$ relativized to a ball in $C^* \times C^*$ containing S is metrizable. It follows that S is weak* compact.

Let us now examine sufficient conditions for membership of S . Take $(\mu, \gamma) \in S$. It is evident from (1.4) that

$$(1.6) \quad \mu \geq 0, \quad \gamma \geq 0, \quad |\mu|_{C^*} \leq T - t_0, \quad \text{and} \quad |\gamma|_{C^*} \leq 1.$$

Here, positivity is understood in the sense that $\langle \mu, h \rangle \geq 0$ when h is nonnegative-valued. $|\cdot|_{C^*}$ denotes the dual norm.

A further property that reflects the fact that (μ, γ) is associated with a relaxed arc (t_1, x, m) issuing from the point $(t_0, x(t_0))$ in (t, x) -space and terminating at some point $(t_1, x(t_1)) \in D$ is

$$(1.7) \quad \langle \mu, \phi_t(t, x) + \phi_x(t, x) \cdot v \rangle = \int \phi \, d\gamma - \phi(t_0, x_0)$$

for all functions $\phi \in C^1(\mathbb{R}^{1+n})$. (In this relationship, (t, x, v) is to be regarded as a generic point in $D \times \Omega$.)

To see this, we calculate

$$\begin{aligned} \langle \mu, \phi_t + \phi_x \cdot v \rangle &= \int_{t_0}^{t_1} \left[\frac{\partial}{\partial t} \phi(t, x(t)) + \frac{\partial}{\partial x} \phi(t, x(t)) \cdot \int v \, dm_t(v) \right] dt \\ &= \int_{t_0}^{t_1} \left[\frac{\partial}{\partial t} \phi(t, x(t)) + \frac{\partial}{\partial x} \phi(t, x(t)) \cdot \dot{x}(t) \right] dt = \int_{t_0}^{t_1} \frac{d}{dt} \phi(t, x(t)) \, dt \\ &= \phi(t_1, x(t_1)) - \phi(t_0, x_0) = \int \phi \, d\gamma - \phi(t_0, x_0). \end{aligned}$$

We use these conditions to define a new set of linear functionals \mathcal{W} . In this case, the chosen symbol indicates that these functionals are associated with relaxed arcs for (C) in some “weak” sense as follows:

$$(1.8) \quad \mathcal{W} := \{(\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D) : (\mu, \gamma) \text{ satisfies (1.6) and (1.7)}\}.$$

Since the conditions defining membership of \mathcal{W} are necessary conditions for membership of S , we have $S \subset \mathcal{W}$. We note also that \mathcal{W} is a convex weak* compact subset of $C^* \times C^*$. Convexity is obvious. \mathcal{W} is weak* compact because it is bounded with respect to the dual norm and expressible as the intersection of a family of weak* closed subspaces and half-spaces.

Our preliminary findings concerning the two subsets S and \mathcal{W} of $C^*(D \times \Omega) \times C^*(D)$, and their interrelation may be summarized as follows: *S and \mathcal{W} are nonempty weak* compact subsets of $C^*(D \times \Omega) \times C^*(D)$, S is convex, and $S \subset \mathcal{W}$.*

1.3. The structure of generalized flows. Our goal is to derive optimality conditions for minimizing arcs by analysing the properties of the set \mathcal{W} of generalized flows. The feasibility of this approach hinges, of course, on our ability to establish a close relationship between the set S of relaxed arcs and the set \mathcal{W} . The two sets do not coincide. Indeed, if ν_1 and ν_2 are two distinct elements in S , the point $\nu \in C^* \times C^*$ defined by

$$\langle \nu, g \rangle = \frac{1}{2} \langle \nu_1, g \rangle + \frac{1}{2} \langle \nu_2, g \rangle$$

(for all $g \in C \times C$) defines an element in \mathcal{W} but not in S . It is the case, however, that the two sets are related through the operation of “convex closure.” The following theorem, which discusses the details of this relationship, has a crucial role in the theory.

THEOREM 1.3. *Let the sets $S, \mathcal{W} \subset C^*(D \times \Omega) \times C^*(D)$ be as defined in (1.5) and (1.8). Then \mathcal{W} is the weak* closed convex hull of S , i.e., \mathcal{W} is the intersection of all weak* closed convex sets containing S .*

A proof of this theorem is given in §1.4. It is similar to one of Fleming [3] and is, in essence, a specialization of arguments in [4].

As a corollary, we have that an arbitrary element in \mathcal{W} can be decomposed into a convex combination of elements in S ; the weightings assigned to individual elements in S are specified through a probability measure on the Borel sets of S . (Here we refer to the Borel sets generated by the dual norm topology on S , which coincide with those generated by the weak* topology.)

COROLLARY 1.4. *Let ν be an arbitrary element in \mathcal{W} . Then there exists a regular probability measure Λ on the Borel subsets of S such that*

$$(1.9) \quad \langle \nu, g \rangle = \int_S \langle \lambda, g \rangle d\Lambda(\lambda)$$

for all $g \in C(D \times \Omega) \times C(D)$.

Proof. Denote by Π the class of regular probability measures on S . For any $\Lambda \in \Pi$ and $g \in C(D \times \Omega) \times C(D)$, the function $\lambda \rightarrow \langle \lambda, g \rangle$ is weak* continuous on S and therefore Λ -integrable. It follows that the right side of (1.9) is well defined. For each $\Lambda \in \Pi$, the mapping

$$g \rightarrow \int_S \langle \gamma, g \rangle d\Lambda(\gamma)$$

is a bounded linear functional on $C(D \times \Omega) \times C(D)$ and hence defines an element in the dual space $C^* \times C^*$. Denote by \tilde{S} the subset of points in $C^* \times C^*$ generated in this way by allowing Λ to range over the set Π . It is straightforward to show that \tilde{S} is a convex set containing S . The set \tilde{S} is also weak* compact. To see this, take a weak* convergent sequence $\{\nu_i\}$ of measures with limit ν , expressible in terms of measures Λ_i , $i = 1, 2, \dots$ drawn from the set Π according to

$$\langle \mu_i, g \rangle = \int_S \langle \gamma, g \rangle d\Lambda_i(\gamma) \quad \text{for all } g \in C(D \times \Omega) \times C(D).$$

By limiting attention to a subsequence, if necessary, we can arrange that the Λ_i 's (or, strictly speaking, the bounded linear functionals associated with the Λ_i 's) converge weak* to some $\Lambda \in \Pi$. Passage to the limit now gives (1.9) for all $g \in C(D \times \Omega) \times C(D)$. In other words, $\mu \in \tilde{S}$. \tilde{S} is therefore sequentially weak* compact. Since it is bounded in the dual norm, it is weak* compact. Then, however, \tilde{S} must contain \mathcal{W} , which, according to Theorem 1.3, lies in the intersection of all weak* closed convex sets containing S . This is what the corollary asserts. \square

A minor refinement of this last result will be required. Recall that any positive bounded linear functional $\nu = (\mu, \gamma)$ on $C(D \times \Omega) \times C(D)$ has associated with it an element $(d\mu, d\gamma)$ comprising bounded, regular positive Borel measures $d\mu$ and $d\gamma$ on $D \times \Omega$ and D , respectively. These measures provide an extension of the linear functional ν , namely,

$$g \rightarrow \int_{D \times \Omega} g_1 d\mu + \int_D g_2 d\gamma,$$

to the class of “integrands” $g = (g_1, g_2)$ in which g_1 and g_2 are extended-value functions on $D \times \Omega$ and D , respectively, which are Borel measurable and bounded below. Of particular interest to us will be integrands in this class comprising pairs of lower semicontinuous functions.

COROLLARY 1.5. *Take any $\nu \in \mathcal{W}$ and let Λ be the measure of Corollary 1.4. Then, for any element $g = (g_1, g_2)$ comprising functions $g_1: D \times \Omega \rightarrow \mathbb{R} \cup \{+\infty\}$ and $g_2: D \rightarrow \mathbb{R} \cup \{+\infty\}$, which are lower semicontinuous functions, (1.9) remains valid when we interpret the functional evaluations $\langle \nu, g \rangle$ and $\langle \lambda, g \rangle$ in the extended sense described above.*

Proof. Let μ , Λ and $g = (g_1, g_2)$ be as in the statement of the corollary. Let $\{g_1^i\}$ and $\{g_2^i\}$ be sequences of uniformly bounded, continuous functions, minorizing g_1 and g_2 and converging everywhere to g_1 and g_2 , respectively. For $i = 1, 2, \dots$, (1.9) is valid when (g_1^i, g_2^i) is substituted for (g_1, g_2) . Such sequences exist [1, p. 212]. Corollary 1.4 is now proved by expressing the functional operations in terms of measures and passage to the limit with the help of the monotone convergence theorem. \square

1.4. Proof of Theorem 1.3. Suppose that the assertion is false. We may then choose an element $(\mu_0, \gamma_0) \in \mathcal{W} \setminus \overline{\text{co}} S$. Here $\overline{\text{co}} S$ denotes the weak* convex closure of S . Therefore, $\{(\mu_0, \gamma_0)\}$ and $\overline{\text{co}} S$ are strictly separated. This means that there exist a function $(l_0, g_0) \in C(D \times \Omega) \times C(D)$ and a number a such that

$$(1.10) \quad \langle \mu, l_0 \rangle + \langle \gamma, g_0 \rangle \geq a \quad \text{for all } (\mu, \gamma) \in S$$

and

$$(1.11) \quad \langle \mu_0, l_0 \rangle + \langle \gamma_0, g_0 \rangle < a.$$

Noting that continuous functions on $D \times \Omega$ and D can be uniformly approximated by Lipschitz continuous functions, we are justified in replacing l_0 and g_0 by new functions that are Lipschitz continuous and that (following possible adjustment of the number a) satisfy (1.10) and (1.11).

Now assume that l_0, g_0 have been extended to the whole of $\mathbb{R}^{1+2n}, \mathbb{R}^{1+n}$, respectively, as Lipschitz functions. (This can be done, for example, by constant extrapolation of values at the boundary of each of the sets $D \times \Omega, D$ along rays radiating from some fixed interior point.)

For each $(t, x) \in (-\infty, T] \times \mathbb{R}^n$, we define

$$(1.12) \quad \phi(t, x) := \inf \left\{ \int_t^{t_1} \int l_0(s, y(s), v) dm_s(v) ds + g_0(t_1, y(t_1)) \right\},$$

where the infimum is taken over elements $\{t_1, y, m\}$, which comprise a point $t_1 \in [t, T]$, a Lipschitz continuous function $y: [t, t_1] \rightarrow \mathbb{R}^n$, and a measurable function $s \rightarrow \mu_s$ mapping the interval $[t, t_1]$ into the space of regular probability measures on Ω , which satisfy

$$\begin{aligned} y(t) &= x, \\ \dot{y}(s) &= \int v dm_s(v) \quad \text{a.e. } s \in [t, t_1]. \end{aligned}$$

Routine modification of standard arguments (see [2]) gives us the following lemma.

LEMMA 1.6. *The function $\phi: (-\infty, t_1] \times \mathbb{R}^n \rightarrow \mathbb{R}$ is locally Lipschitz continuous.*

Inequality (1.10) can be expressed as

$$\int_{t_0}^{t_1} \int l_0(s, y(s), v) dm_s(v) dt + g_0(t_1, y(t_1)) \geq a$$

for all relaxed arcs (t_1, y, m) for system (C). Now, in view of (1.3), the constraint

$$“(s, y(s)) \in D \quad \text{for all } s \in [t, t_1]”$$

is inactive. By definition of ϕ , then,

$$(1.13) \quad \phi(t_0, x_0) \geq a.$$

LEMMA 1.7. *Let (t, x) be a point in $(-\infty, T) \times \mathbb{R}^n$ at which ϕ is differentiable. Then, for all $v \in \Omega$,*

$$\frac{\partial}{\partial t} \phi(t, x(t)) + \frac{\partial}{\partial x} \phi(t, x(t)) \cdot v + l_0(t, x, v) \geq 0.$$

Proof. For each $v \in \Omega$ and $\varepsilon \in (0, T - t)$, the “arc”

$$(t + \varepsilon, y(s) \equiv x + (s - t)v, \mu_s \equiv \delta_{\{v\}})$$

lies in the set of elements over which the infimum in (1.12) is taken. It follows that

$$\varepsilon^{-1} [g_0(t + \varepsilon, x + \varepsilon v) - \phi(t, x)] \geq -\varepsilon^{-1} \int_t^{t+\varepsilon} l_0(s, x + (s - t)v, v) ds.$$

The definition of ϕ also implies that

$$g_0(t + \varepsilon, x + \varepsilon v) \geq \phi(t + \varepsilon, x + \varepsilon v).$$

The assertions of the lemma are now proved by combining these inequalities and passing to the limit as $\varepsilon \downarrow 0$. (This is justified under the differentiability assumption and because the integrand is continuous.) \square

Consider now the mollifiers $\rho_i : (-\infty, 0] \times \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, 2, \dots$ defined according to

$$\rho_i(t, x) = \begin{cases} i^{(n+1)} & \text{if } -i^{-1} \leq t \leq 0 \quad \text{and} \quad -i^{-1} \leq x_i \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Given a locally integrable function h on \mathbb{R}^{1+n} , the function $h * \rho_i$ is assumed to be

$$(h * \rho_i)(t, x) := \iint_{\mathbb{R}^{n+1}} h(t + \tau, x + \xi) \rho_i(\tau, \xi) d\tau d\xi.$$

The following lemma lists properties of $h * \rho_i$.

LEMMA 1.8. *Let $h : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ be a locally Lipschitz continuous function and let ρ_i , $i = 1, 2, \dots$, be as defined above. Then*

- (i) $h * \rho_i \in C^1(\mathbb{R}^{1+n})$ for $i = 1, 2, \dots$;
- (ii) $\nabla(h * \rho_i) = \iint_{\mathbb{R} \times \mathbb{R}^n} \nabla h(t + \tau, x + \xi) \rho_i(\tau, \xi) d\tau d\xi$ for $i = 1, 2, \dots$;
- (iii) $(h * \rho_i)(t, x) \rightarrow h(t, x)$ uniformly on compact sets.

The mollifiers are now used to establish existence of smooth approximate subsolutions to the Hamilton-Jacobi equation.

LEMMA 1.9. *There exist a sequence $\{\phi^i\}$ in $C^1(\mathbb{R}^{n+1})$ and a sequence of numbers $\{\delta_i\}$, with $\delta_i \downarrow 0$, such that $\phi^i(t, x) \rightarrow \phi(t, x)$, uniformly on D , as $i \rightarrow \infty$, and*

$$(1.14) \quad \frac{\partial}{\partial t} \phi^i(t, x) + \frac{\partial}{\partial x} \phi^i(t, x) \cdot v + l_0(t, x, v) \geq -\delta_i \quad \text{for } (t, x) \in D, \quad v \in \Omega,$$

$$i = 1, 2, \dots$$

Proof. Extend ϕ to all of $\mathbb{R} \times \mathbb{R}^n$ as a locally Lipschitz continuous function and define $\phi^i := \phi * \rho_i$ for $i = 1, 2, \dots$. According to Lemma 1.8, each ϕ^i is a C^1 function and $\phi^i \rightarrow \phi$ uniformly on D . By the almost everywhere differentiability of locally Lipschitz functions and by Lemma 1.7, we know that, for all $v \in \Omega$ and almost every $(t, x) \in (-\infty, t_1] \times \mathbb{R}^n$,

$$0 \leq \frac{\partial}{\partial t} \phi(t, x) + \frac{\partial}{\partial x} \phi(t, x) \cdot v + l_0(t, x, v).$$

For each i , the mollifier ρ_i is nonnegative-valued. The inequality is therefore preserved on convolving both sides of the inequality with ρ_i . Writing $\phi^i := \phi * \rho_i$ and appealing to Lemma 1.8, we deduce that

$$0 \leq \phi_t^i(t, x) + \phi_x^i(t, x) \cdot v + (l_0 * \rho_i)(t, x, v) \quad \text{for all } (t, x, v) \in D\mathbf{x}\Omega.$$

Now set

$$\delta_i = |l_0 - l_0 * \rho_i|_{C(D)}, \quad i = 1, 2, \dots$$

According to Lemma 1.8, $\delta_i \downarrow 0$ as $i \rightarrow \infty$. Equation (1.14) is then valid for this choice of $\{\delta_i\}$. \square

We are now ready for completion of the proof of Theorem 1.3. Take $\{\phi^i\}$ and $\{\delta_i\}$ as in Lemma 1.9. By definition of the set \mathcal{W} of linear functionals, the element $(\mu_0, \gamma_0) \in \mathcal{W}$ satisfies

$$\langle \mu_0, \phi_t^i + \phi_x^i \cdot v \rangle = \langle \gamma_0, \phi^i \rangle - \phi^i(t_0, x_0) \quad \text{for } i = 1, 2, \dots$$

Then, by (1.14),

$$\begin{aligned} \phi^i(t_0, x_0) &= \langle \mu_0, -(\phi_t^i + \phi_x^i \cdot v) \rangle + \langle \gamma_0, \phi^i \rangle \\ &\leq \langle \mu_0, l_0 \rangle + \langle \gamma_0, \phi \rangle + \varepsilon_i, \end{aligned}$$

where $\varepsilon_i = \delta_i |\mu_0|_{C^*} + |\phi^i - \phi|_{C(D)}$. By Lemma 1.9, $\varepsilon_i \rightarrow 0$ and $\phi^i(t_0, x_0) \rightarrow \phi(t_0, x_0) \rightarrow \phi(t_0, x_0)$ as $i \rightarrow \infty$. It follows now from (1.11) and (1.13) that

$$a \leq \phi(t_0, x_0) = \lim_i \phi^i(t_0, x_0) \leq \langle \mu_0, l_0 \rangle + \langle \gamma_0, \phi \rangle < a.$$

We conclude from this contradiction that $\mathcal{W} = \overline{\text{co}} S$.

2. Optimality conditions. Our knowledge about generalized flows is now used, in the derivation of optimality conditions. We consider an optimal control problem (P) in which the terminal time t_1 is a choice variable (this is a “free time” problem), where a description of the dynamics is provided in terms of a differential inclusion and where constraints, expressed in terms of general set inclusions, are imposed on values of the state trajectory $x(\cdot)$, the terminal time t_1 , and the terminal in state $x(t_1)$. Problem (P) is as follows:

(P) Minimize $\int_{t_0}^{t_1} l(t, x(t), \dot{x}(t)) dt + g(t_1, x(t_1))$ over points $t_1 \in [t_0, T]$ and Lipschitz continuous functions $x: [t_0, t_1] \rightarrow \mathbb{R}^n$ that satisfy

$$(2.1) \quad \dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } t \in [t_0, t_1],$$

$$(2.2) \quad (t, x(t)) \in A \quad \text{for all } t \in [t_0, t_1],$$

$$(2.3) \quad x(t_0) = x_0, \quad (t_1, x(t_1)) \in C.$$

In problem (P), the interval $[t_0, T]$, the point $x_0 \in \mathbb{R}^n$, the extended-value functions $l: [t_0, T] \times \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ and $g: [t_0, T] \times \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$, the multifunction $F: [t_0, T] \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, and the set $A \subset [t_0, T] \times \mathbb{R}^n$ are given. We refer to elements (t_1, x) , in which $t_1 \in [t_0, T]$ and x is a Lipschitz continuous \mathbb{R}^n -valued function on $[t_0, t_1]$ satisfying (2.1)–(2.3), as *admissible trajectories* for problem (P).

The data for problem (P) satisfies the following hypotheses:

(H1) l and g are lower semicontinuous;

(H2) F takes values compact, convex sets, and the set

$$\{(v, (t, x)) \in \mathbb{R}^n \times ([t_0, T] \times \mathbb{R}^n): v \in F(t, x), (t, x) \in A\}$$

is compact;

(H3) A and C are compact sets;

(H4) There exists an admissible trajectory (\bar{t}_1, \bar{x}) for problem (P) for which

$$\int_{t_0}^{\bar{t}_1} l(t, \bar{x}(t), \dot{\bar{x}}(t)) dt + g(\bar{t}_1, \bar{x}(\bar{t}_1)) < \infty;$$

(H5) For each $(t, x) \in [t_0, T] \times \mathbb{R}^n$, the function $l(t, x, \cdot)$ restricted to the convex set $F(t, x)$ is a convex function.

2.1. The main results. We shall show in § 2.4 that our earlier analysis of the structure of generalized flows leads to the following dual representation of the minimum cost for problem (P).

THEOREM 2.1. *Under hypotheses (H1)–(H5), problem (P) has a solution. The infimum cost, $\inf P$, is*

$$\inf (P) = \sup \{ \phi(t_0, x_0) \},$$

where the supremum on the right is taken over functions ϕ in $C^1(\mathbb{R}^{1+n})$ that satisfy

$$\frac{\partial}{\partial t} \phi(t, x) + \frac{\partial}{\partial x} \phi(t, x) \cdot v + l(t, x, v) \geq 0 \quad \text{for all } (t, x) \in A, \quad v \in F(t, x)$$

and

$$\phi(t, x) \leq g(t, x) \quad \text{for all } (t, x) \in C.$$

The above theorem leads directly to the following optimality conditions.

THEOREM 2.2. *Let $(t_1^*, x^*(\cdot))$ be an admissible trajectory for problem (P). Assume that hypotheses (H1)–(H5) are satisfied. Then the following conditions hold.*

(i) Sufficient condition. Suppose that there exists a sequence of functions $\{\phi^i\}$ in $C^1(\mathbb{R}^{1+n})$ such that, for $i = 1, 2, \dots$,

$$(2.4) \quad \frac{\partial}{\partial t} \phi^i(t, x) + \frac{\partial}{\partial x} \phi^i(t, x) \cdot v + l(t, x, v) \geq 0 \quad \text{for all } (t, x) \in A, \quad v \in F(t, x),$$

$$(2.5) \quad \phi^i(t, x) \leq g(t, x) \quad \text{for all } (t, x) \in C,$$

and furthermore

$$(2.6) \quad \lim_i \phi^i(t_0, x_0) = \int_{t_0}^{t_1^*} l(t, x^*(t), \dot{x}^*(t)) dt + g(t_1^*, x^*(t_1)).$$

Then $(t_1^*, x^*(\cdot))$ is a minimizer.

(ii) Necessary condition. Suppose that $(t_1^*, x^*(\cdot))$ is a minimizer. Then there exists a sequence $\{\phi^i\}$ in $C^1(\mathbb{R}^{1+n})$ satisfying conditions (2.4)–(2.6).

Proof. (i) Let $\{\phi_i\}$ be a sequence with the stated properties and let $(t_1, x(\cdot))$ be an arbitrary admissible trajectory for (P). Since, for each i , the function $t \rightarrow \phi_i(t, x(t))$ is Lipschitz continuous, we may express it as the integral of its derivative. Thus

$$\phi_i(t_0, x_0) = - \int_{t_0}^{t_1} \frac{d}{dt} \phi_i(t, x(t)) dt + \phi_i(t_1, x(t_1)).$$

Using the chain rule to reduce the integrand and noting the boundary condition (2.5) on ϕ_i , we obtain

$$\phi_i(t_0, x_0) \leq - \int_{t_0}^{t_1} \left\{ \frac{\partial}{\partial t} \phi_i(t, x(t)) + \frac{\partial}{\partial x} \phi_i(t, x(t)) \cdot \dot{x}(t) \right\} dt + g(t_1, x(t_1)).$$

Since ϕ satisfies (2.4), we conclude that

$$\phi_i(t_0, x_0) \leq \int_{t_0}^{t_1} l(t, x(t), \dot{x}(t)) dt + g(t_1, x(t_1)).$$

In the limit $i \rightarrow \infty$, we obtain

$$\lim_{i \rightarrow \infty} \phi_i(t_0, x_0) \leq \int_{t_0}^{t_1} l(t, x(t), \dot{x}(t)) dt + g(t_1, x(t_1)).$$

This inequality combines with (2.6) to establish that $(t_1^*, x^*(\cdot))$ is a minimizer.

(ii) Now suppose that $(t_1^*, x^*(\cdot))$ is a minimizer. Let $\{\phi_i\}$ be a maximizing sequence associated with the supremum in the statement of Theorem 2.1. Then the ϕ_i 's automatically satisfy condition (2.4) and (2.5). The remaining condition (2.6) is also satisfied, since, by Theorem 2.1, the supremum coincides with the infimum cost for (P). \square

In problem (P), the model for the system dynamics is a differential inclusion. There is, however, considerable flexibility built into the formulation, since the cost integrand l is allowed to be a lower semicontinuous extended-valued function. Suppose, for example, that the differential inclusion was replaced by a differential equation with control as follows:

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [t_0, t_1], \quad u(t) \in U,$$

and the integral functional term in the cost were of the form $\int_{t_0}^{t_1} l'(t, x(t), u(t)) dt$. Minimization is now conducted over Lipschitz continuous functions x and measurable functions u satisfying the constraints. The problem assumes the guise of (P) when we choose the multifunction F and the extended-value cost integrand l to be

$$F(t, x) := f(t, x, U) \quad \text{and} \quad l(t, x, v) := \inf \{l'(t, x, u) : v = f(t, x, u)\},$$

respectively. (We interpret the infimum as $+\infty$ at points for which $v \notin f(t, x, U)$.) Under mild hypotheses, we arrive at an equivalent problem in this way (see [6]). The conditions on F and l in (H1) and (H2) are satisfied if, for example, l' and f are continuous functions, U is a compact set, and $f(t, x, U)$ is convex. Theorem 2.2 then provides optimality conditions for optimal control problems involving differential equations with control, via the reformulation.

2.2. A weak formulation. Our object here is to reformulate problem (P) as an optimization problem over a space of linear functionals. Choose positive numbers r , K such that

$$\begin{aligned} |v| &\leq r \quad \text{for all } v \in F(t, x), \quad (t, x) \in A, \\ C, A &\subset [t_0, T] \times (K\bar{B}), \\ (2.7) \quad |x_0| + |T - t_0|r &\leq K. \end{aligned}$$

Such choices are possible in view of hypotheses (H2) and (H3). We write $D = [t_0, T] \times (K\bar{B})$ and $\Omega = r\bar{B}$.

The optimization problem here of interest, denoted by (W), is as follows:

(W) Minimize $\int_{D \times \Omega} l d\mu + \int_D g d\gamma$ over elements $(\mu, \gamma) \in \mathcal{W}$ that satisfy

$$\langle \mu, d_{\tilde{F}(t,x)}(v) \rangle = 0, \quad \langle \mu, d_A(t, x) \rangle = 0, \quad \langle \gamma, d_C(t, x) \rangle = 0.$$

In these relationships, \tilde{F} is the multifunction

$$\tilde{F}(t, x) := \begin{cases} F(t, x) & \text{if } (t, x) \in A, \\ \mathbb{R}^n & \text{if } (t, x) \notin A, \end{cases}$$

and $d_Z(\cdot): \mathbb{R}^k \rightarrow \mathbb{R}$ denotes the Euclidean distance function associated with a set $Z \subset \mathbb{R}^k$. Generic elements in $C(D \times \Omega)$ and $C(D)$ are written as (t, x, v) and (t, x) , respectively.

Problem (W) requires some interpretation. Recall that \mathcal{W} is the subset of $C^*(D \times \Omega) \times C^*(D)$ comprising elements that satisfy

$$\mu \geq 0, \quad \gamma \geq 0, \quad |\mu|_{C^*} \leq T - t_0, \quad |\gamma|_{C^*} \leq 1,$$

and

$$\left\langle \mu, \left[\frac{\partial}{\partial t} \phi(t, x) + \frac{\partial}{\partial x} \phi(t, x) \cdot v \right] \right\rangle = \langle \gamma, \phi \rangle - \phi(t_0, x_0)$$

for every $\phi \in C^1(\mathbb{R}^{1+n})$.

Given $(\mu, \gamma) \in \mathcal{W}$, the cost function involves integration with respect to the regular Borel measures on $D \times \Omega$ and D , which represent μ and γ , respectively. The integrals are well defined, since the integrands l and g are lower semicontinuous (and therefore Borel measurable and bounded below on the compact domains $D \times \Omega$ and D).

It is easy to see that

$$(2.8) \quad \inf(P) \geq \inf(W).$$

This follows from the fact that, if the arc x satisfies the constraints of problem (P), then the element $(\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D)$ defined by

$$\langle \mu, \xi \rangle := \int_{t_0}^{t_1} \xi(t, x(t), \dot{x}(t)) dt \quad \text{and} \quad \langle \gamma, \eta \rangle = \eta(t_1, x(t_1))$$

for all $\xi \in C(D \times \Omega)$ and $\eta \in C(D)$ satisfies the constraints of problem (W), and the values of the costs are the same. We now use the machinery developed in §1 to obtain the converse inequality.

THEOREM 2.3. *Problems (P) and (W) have minimizers, and $\inf(P) = \inf(W)$.*

Proof. First, note that the set of elements satisfying the constraints in (W) is a weak* closed subset of the weak* compact set \mathcal{W} and, as such, is weak* compact. It is nonempty, since it contains the element associated with the admissible arc (\bar{t}, \bar{x}) of hypothesis (H4). The cost functional for (W) is weak* lower semicontinuous on $\{(\mu, \gamma) \in C^* \times C^*: \mu \geq 0, \gamma \geq 0\}$. To verify this, we must show that, for any $\alpha \in \mathbb{R}$, the set

$$S_\alpha = \left\{ (\mu, \gamma) \in C^* \times C^*: \int_{D \times \Omega} l d\mu + \int_D g d\gamma \leq \alpha \right\}$$

is weak* closed. It is easily deduced from the monotone convergence theorem that

$$S_\alpha = \bigcap_{i=1}^{\infty} S_\alpha^i,$$

where

$$S_\alpha^i := \left\{ (\mu, \gamma) \in C^* \times C^*: \int_{D \times \Omega} l_i d\mu + \int_D g_i d\gamma \leq \alpha \right\}.$$

In these definitions, $\{l_i\}$ is some sequence of continuous functions on $D \times \Omega$ minorizing l and converging everywhere to l , and $\{g_i\}$ is any sequence of continuous functions on D minorizing g and converging everywhere to g . The fact that S_α is weak* closed now follows from the weak* closedness of the S_α^i 's. Existence of a minimizer (μ_0, γ_0) is therefore assured.

We now demonstrate the existence of an admissible arc for (P), the value of whose cost coincides with that of (μ_0, γ_0) in problem (W); this fact combined with (2.8) will permit us to conclude that (P), too, has a minimizer and that $\inf(P) = \inf(W)$.

Since $(\mu_0, \gamma_0) \in \mathcal{W}$, we know from Corollary 1.4 of the existence of a regular Borel probability measure Λ on S such that

$$(2.9) \quad (\mu_0, \gamma_0) = \int_S (\mu, \gamma) d\Lambda(\mu, \gamma).$$

The fact that (μ_0, γ_0) is feasible for problem (W) implies that

$$\begin{aligned} \int_S \langle \mu, d_{F(t,x)}(v) \rangle d\Lambda &= 0, \\ \int_S \langle \mu, d_A(t, x) \rangle d\Lambda &= 0, \quad \text{and} \\ \int_S \langle \gamma, d_C(t, x) \rangle d\Lambda &= 0. \end{aligned}$$

Let S' be the weak* closed subset of points (μ, γ) in S such that

$$(2.10) \quad \langle \mu, d_{F(t,x)}(v) \rangle = \langle \mu, d_A(t, x) \rangle = \langle \gamma, d_C(x) \rangle = 0.$$

Since the integrands involved are nonnegative, we deduce that S' has full Λ -measure.

Turn now to consideration of the cost function for (W). Equation (2.9) and the fact that S' has full measure imply that

$$(2.11) \quad \int l d\mu_0 + \int g d\gamma_0 = \int_{S'} \left[\int l d\mu + \int g d\gamma \right] d\Lambda(\mu, \gamma).$$

Suppose that $\int l d\mu + \int g d\gamma > \int l d\mu_0 + \int g d\gamma_0$ for every $(\mu, \gamma) \in S'$. Then, since Λ restricted to the Borel subsets for S' is a probability measure, we have

$$\int_{S'} \left[\int l d\mu + \int g d\gamma \right] d\Lambda(\mu, \gamma) > \int l d\mu_0 + \int g d\gamma_0.$$

This contradicts (2.11). Hence there exists some element (μ, γ) in S' such that

$$\int l d\mu + \int g d\gamma = \int l d\mu_0 + \int g d\gamma_0.$$

Let (t_1, x, m) be the relaxed arc associated with (μ, γ) . We have

$$(2.12) \quad \int_{t_0}^{t_1} \int l(t, x(t), v) dm_t(v) dt + g(t_1, x(t_1)) = \int l d\mu_0 + \int g d\gamma_0.$$

It follows from (2.10) that

$$(2.13) \quad \int d_{\tilde{F}(t, x(t))}(v) dm_t(v) = 0 \quad \text{a.e. } [t_0, t_1],$$

$$(2.14) \quad d_A(t, x(t)) = 0 \quad \text{a.e. } [t_0, t_1],$$

and

$$(2.15) \quad d_C(t_1, x(t_1)) = 0.$$

Equations (2.14) and (2.15) show us that

$$(2.16) \quad \text{graph } \{x\} \subset A \quad \text{and} \quad (t_1, x(t_1)) \in C.$$

From (2.13) and (2.14) we deduce that

$$(2.17) \quad \text{support } \{m_t\} \subset F(t, x(t)) \quad \text{a.e.}$$

Define

$$(2.18) \quad c(t) := \int_{t_0}^t \left(\int l(t, x(t), v) dm_t(v) \right) dt.$$

Then

$$\begin{bmatrix} \dot{c}(t) \\ \dot{x}(t) \end{bmatrix} = \int \begin{bmatrix} l(t, x(t), v) \\ v \end{bmatrix} dm_t(v) \quad \text{a.e.}$$

This relationship, together with (2.17), imply that, for almost every t ,

$$\begin{aligned} \begin{bmatrix} \dot{c}(t) \\ \dot{x}(t) \end{bmatrix} &\in \overline{\text{co}} \left\{ \begin{bmatrix} l(t, x(t), v) \\ v \end{bmatrix} : v \in F(t, x(t)) \right\} \\ &\subset \{(\alpha, v) \in \mathbb{R} \times \mathbb{R}^n : \alpha \geq l(t, x(t), v) \text{ and } v \in F(t, x(t))\} \\ &\quad \text{a.e. } [t_0, t_1]. \end{aligned}$$

This follows from the fact that the restriction of $l(t, x(t), \cdot)$ to the convex set $F(t, x(t))$ is a lower semicontinuous, convex function. Then, however,

$$(2.19) \quad \dot{c}(t) \geq l(t, x(t), \dot{x}(t)) \quad \text{a.e. } [t_0, t_1]$$

and

$$(2.20) \quad \dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } [t_0, t_1].$$

From (2.12), (2.18), and (2.19), we deduce that

$$(2.21) \quad \int l d\mu_0 + \int g d\gamma_0 \geq \int_{t_0}^{t_1} l(t, x(t), \dot{x}(t)) dt + g(t_1, x(t_1)).$$

In view of (2.16) and (2.20), (t_1, x) is an admissible arc for (P), while (2.21) establishes that this arc has cost not greater than $\inf(W)$. The proof is complete. \square

2.3. Fenchel duality. This section provides a brief summary of aspects of convex analysis relevant to the analysis of the weak problem. It focusses on one procedure (Fenchel duality) for generating optimization problems “paired in duality.” For a fuller account, see [10].

Take \mathcal{Y} and \mathcal{Y}' to be locally convex topological vector spaces, in duality with respect to a bilinear form $\langle \cdot, \cdot \rangle : \mathcal{Y} \times \mathcal{Y}' \rightarrow \mathbb{R}$. This means that every continuous linear functional on \mathcal{Y} can be represented by $\langle \cdot, y' \rangle$ for some $y' \in \mathcal{Y}'$ and that every continuous linear functional on \mathcal{Y}' can be represented by $\langle y, \cdot \rangle$ for some $y \in \mathcal{Y}$.

DEFINITION 2.4. (i) Consider a function $p : \mathcal{Y} \rightarrow \bar{\mathbb{R}}$. The *convex conjugate* of p is the function $p' : \mathcal{Y}' \rightarrow \bar{\mathbb{R}}$, defined by

$$(2.22) \quad p'(y') := \sup \{ \langle y, y' \rangle - p(y) : y \in \mathcal{Y} \}.$$

(ii) Consider a function $q : \mathcal{Y} \rightarrow \bar{\mathbb{R}}$. The *concave conjugate* of q is the function $q' : \mathcal{Y}' \rightarrow \bar{\mathbb{R}}$, defined by

$$(2.23) \quad q'(y') := \inf \{ \langle y, y' \rangle - q(y) : y \in \mathcal{Y} \}.$$

It will always be evident from context whether a convex or concave conjugate is intended, and confusion should not arise from using the prime notation for both constructs.

Because the spaces \mathcal{Y} and \mathcal{Y}' are interchangeable, we can apply the operation of convex conjugation to functions on \mathcal{Y}' , thereby obtaining functions on \mathcal{Y} . In particular, if we start with a function $p: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$, we can construct the convex conjugate $p': \mathcal{Y}' \rightarrow \bar{\mathbb{R}}$ according to (2.22) and then the *second convex conjugate* $p'': \mathcal{Y} \rightarrow \bar{\mathbb{R}}$, namely,

$$p''(y) := \sup \{ \langle y, y' \rangle - p'(y') : y' \in \mathcal{Y}' \}.$$

Likewise, we can define the *second convex conjugate* $q'': \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ of a function $q: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$, namely,

$$q''(y) := \inf \{ \langle y, y' \rangle - q'(y') : y' \in \mathcal{Y}' \},$$

where q' is given by (2.23).

Of particular interest are the conjugates of *proper* functions, as shown in Definition 2.5.

DEFINITION 2.5. (i) A function $p: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ is said to be a *proper convex function* if p is a lower semicontinuous, convex function such that $p(y) > -\infty$ for all points $y \in \mathcal{Y}$ and $p(\bar{y}) < +\infty$ for some point $\bar{y} \in \mathcal{Y}$.

(ii) A function $q: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ is said to be a *proper concave function* if q is an upper semicontinuous, concave function such that $q(y) < +\infty$ for all points $y \in \mathcal{Y}$ and $q(\bar{y}) > -\infty$ for some point $\bar{y} \in \mathcal{Y}$.

In fact, there is a one-to-one relationship between proper convex functions and their convex conjugates; a proper convex function is recoverable from information about its convex conjugate by taking the second convex conjugate. Analogous results apply to proper concave functions.

PROPOSITION 2.6. (i) If a function $p: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ is a *proper convex function*, then its *convex conjugate* is a *proper convex function*. We have

$$p(y) = p''(y) \quad \text{for all } y \in \mathcal{Y},$$

where p'' is the *second convex conjugate* of p .

(ii) If a function $q: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ is a *proper concave function*, then its *concave conjugate* is a *proper concave function*. We have

$$q(y) = q''(y) \quad \text{for all } y \in \mathcal{Y},$$

where q'' is the *second concave conjugate* of q .

A Fenchel program is a convex optimization problem taking the form

$$(I) \quad \text{Minimize } p(y) - q(y) \quad \text{over } y \in \mathcal{Y},$$

in which $p: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ and $q: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ are proper convex and proper concave functions, respectively. We call this the *primal problem*.

Fenchel duality is concerned with relating it to the associated concave problem, the *dual problem*, namely,

$$(II) \quad \text{Maximize } q'(y') - p'(y') \quad \text{over } y' \in \mathcal{Y}',$$

in which p' and q' are the convex and concave conjugates of p and q , respectively.

By the definitions of the conjugate functionals, we have that, for any $y \in \mathcal{Y}$ and $y' \in \mathcal{Y}'$,

$$p(y) - q(y) = \{ \langle y, y' \rangle - q(y) \} - \{ \langle y, y' \rangle - p(y) \} \geq q'(y') - p'(y').$$

It follows that

$$\inf \{ p(y) - q(y) : y \in \mathcal{Y} \} \geq \sup \{ q'(y') - p'(y') : y' \in \mathcal{Y}' \}.$$

Interest centers on circumstances when inequality here can be replaced by equality; in such circumstances, the primal and dual problems are said to be in *strong duality*.

We use the following criterion for strong duality. The hypotheses involved also ensure that the primal problem has a solution.

PROPOSITION 2.7. Let $p: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ and $q: \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ be proper convex and proper concave functions, respectively. Denote by p' and q' the convex and concave conjugate functions of p and q , respectively. We suppose that the following condition holds:

(H) There exists a point $y' \in \mathcal{Y}'$ such that p' is continuous in a neighborhood of y' and $q'(y') > -\infty$.

Then

$$(2.24) \quad \inf \{p(y) - q(y) : y \in \mathcal{Y}\} = \sup \{q'(y') - p'(y') : y' \in \mathcal{Y}'\}.$$

Furthermore, the infimum on the left is achieved. This is true even if the infimum is $+\infty$, in which case, any $y \in \mathcal{Y}$ is interpreted as a minimizer.

2.4. Proof of Theorem 2.1. The proof of Theorem 2.1 involves reformulating the weak problem (W) of §2.2 as a Fenchel program, namely,

$$\text{Minimize } p(y) - q(y) \quad \text{over } y \in \mathcal{Y}$$

and applying the theory of the preceding section.

In this application, the space \mathcal{Y} is chosen to be the space $C^*(D \times \Omega) \times C^*(D)$ with the product weak* topology, and \mathcal{Y}' is chosen to be the space $C(D \times \Omega) \times C(D)$ with the supremum norm topology. The bilinear form pairing these two spaces in duality is simply

$$\langle (\mu, \gamma), (\xi, \eta) \rangle := \langle \mu, \xi \rangle + \langle \gamma, \eta \rangle$$

for $(\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D)$ and $(\xi, \eta) \in C(D \times \Omega) \times C(D)$. The two terms on the right are intended to describe the action of bounded linear functionals μ and γ in the dual spaces $C^*(D \times \Omega)$ and $C^*(D)$ on points ξ and η in $C(D \times \Omega)$ and $C(D)$, respectively.

The reformulation (we use the label (W)') amounts to retaining the value of the cost function on points satisfying the constraints of problem (W) as originally formulated and replacing it by $+\infty$ on points where they are violated. It is as follows:

$$(W)' \quad \text{Minimize } p(\mu, \gamma) - q(\mu, \gamma) \quad \text{over } (\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D).$$

Here $p: C^* \times C^* \rightarrow \mathbb{R} \cup \{+\infty\}$ and $q: C^* \times C^* \rightarrow \mathbb{R} \cup \{-\infty\}$ are defined as follows:

$$p(\mu, \gamma) := \int_{D \times \Omega} l \, d\mu + \int_D g \, d\gamma + \chi_P(\mu, \gamma) \quad \text{and} \quad q(\mu, \gamma) := -\chi_Q(\mu, \gamma).$$

In these expressions, χ_A denotes the indicator function of a set A , below:

$$\chi_A(a) := \begin{cases} +\infty & \text{if } a \notin A, \\ 0 & \text{if } a \in A. \end{cases}$$

The sets P and Q are taken as

$$P := \{(\mu, \gamma) \in C^* \times C^* : \mu \geq 0, \gamma \geq 0, |\mu|_{C^*} \leq (T - t_0), |\gamma|_{C^*} \leq 1,$$

and (μ, γ) satisfies condition (i) below\}

and

$$Q := \{(\mu, \gamma) \in C^* \times C^* : (\mu, \gamma) \text{ satisfies condition (ii) below}\}:$$

- (i) $\langle \mu, d_{\tilde{F}(t, x)}(v) \rangle = 0$, $\langle \mu, d_A(t, x) \rangle = 0$, and $\langle \gamma, d_C(t, x) \rangle = 0$;
- (ii) $\langle \mu, \phi_t + \phi_x \cdot v \rangle = \langle \gamma, \phi \rangle - \phi(t_0, x_0)$ for all $\phi \in C^1(\mathbb{R}^{1+n})$.

The convex and concave conjugates of p and q are

$$(2.25) \quad p'(\xi, \eta) = \sup \{ \langle \mu, \xi \rangle + \langle \gamma, \eta \rangle - p(\mu, \gamma) : (\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D) \}$$

and

$$(2.26) \quad q'(\xi, \eta) = \inf \{ \langle \mu, \xi \rangle + \langle \gamma, \eta \rangle - q(\mu, \gamma) : (\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D) \},$$

respectively.

The information we require about p and q , and their conjugates, is embodied in the following proposition. Here we use the notation b^+ , below, to describe the “non-negative” part of a real-valued function b :

$$b^+(s) := \max \{0, b(s)\}.$$

LEMMA 2.8. *The functions p and q are proper convex and proper concave functions, respectively, with respect to the weak* topology. The convex and concave conjugates p' and q' of these functions are, respectively,*

$$p'(\xi, \eta) = \max_{(t, x, v) \in A \times \text{graph}\{F\}} (\xi - l)^+(t, x, v) \cdot (T - t_0) + \max_{(t, x) \in C} (\eta - g)^+(t, x)$$

and

$$q'(\xi, \eta) = \eta(t_0, x_0) - \chi_{\bar{\mathcal{T}}}(\mu, \gamma),$$

in which $\bar{\mathcal{T}}$ is the closure in the supremum norm topology of the set

$$\begin{aligned} \mathcal{T} &:= \{(\xi', \eta') \in C \times C : \xi'(t, x, v) \\ &= \phi_t(t, x) + \phi_x(t, x) \cdot v, \eta' = -\phi \text{ for some } \phi \in C^1(\mathbb{R}^{1+n})\}. \end{aligned}$$

Proof. The function p can be expressed as $p(\mu, \gamma) = r(\mu, \gamma) + \chi_P(\mu, \gamma)$, where

$$r(\mu, \gamma) := \int_{D \times \Omega} l \, d\mu + \int_D g \, d\gamma + \chi_{\{(\mu', \gamma') \in C^* \times C^* : \mu' \geq 0, \gamma' \geq 0\}}(\mu, \gamma).$$

The function r is finite-valued at the point associated with the admissible trajectory (\bar{l}, \bar{x}) of hypothesis (H3). r cannot take the value $-\infty$, since l and g are bounded below, and r is lower semicontinuous according to our earlier observations. r then is a proper convex function. The set P is expressible as the intersection of a family of weak* closed sets of the following form:

$$\{(\mu, \gamma) \in C^*(D \times \Omega) \times C^*(D) : \langle (\mu, \gamma), d \rangle \leq \alpha\}$$

for some $d \in C(D \times \Omega) \times C(D)$ and $\alpha \in \mathbb{R}$. P is nonempty since it contains the element $(\bar{\mu}, \bar{\gamma})$ corresponding to the admissible trajectory (\bar{l}_1, \bar{x}) . As such, P is a nonempty, convex, weak* closed set. Along similar lines, we show that the set Q is also nonempty, convex, and weak* closed. It follows that p and $-q$ are proper convex functions. (The indicator functions of nonempty, closed, convex sets are proper convex functions, and the class of lower semicontinuous, convex functions is closed under summation.) Then q ($= -(-q)$) also is a proper concave function.

We now compute the values of p' . Take any $(\xi, \eta) \in C(D \times \Omega) \times C(D)$. Evidently,

$$(2.27) \quad p'(\xi, \eta) = \sup \left\{ \int_{D \times \Omega} (\xi - l) \, d\mu : \mu \in P_1 \right\} + \sup \left\{ \int_D (\eta - g) \, d\gamma : \gamma \in P_2 \right\},$$

in which

$$\begin{aligned} P_1 &:= \left\{ \mu \in C^*(D \times \Omega) : \mu \geq 0, \int_{D \times \Omega} d_{F(t, x)}(v) \, d\mu = 0, \right. \\ &\quad \left. \int_{D \times \Omega} d_A(t, x) \, d\mu = 0 \text{ and } \int_{D \times \Omega} d\mu \leq (T - t_0) \right\} \end{aligned}$$

and

$$P_2 := \left\{ \gamma \in C^*(D): \gamma \geq 0, \int_D d_C d\gamma = 0 \quad \text{and} \quad \int_D d\gamma \leq 1 \right\}.$$

Consider the first supremum in (2.27). It is clear from the conditions defining set P_1 that the support of each μ in P_1 is contained in $\text{graph}\{F\} \cap A$. If $\eta - g$ is nonpositive-valued, the supremum is zero. Otherwise, the supremum is achieved by concentrating the measure at a point in the compact set $\text{graph}\{F\} \cap A$ where the upper semicontinuous function $\eta - g$ takes its maximum value. The supremum is

$$\max_{A \times \text{graph}\{F\}} (\xi - l)^+ \cdot (T - t_0).$$

The value of the second supremum in (2.9) is likewise shown to be $\max_C (\eta - g)^+$. This verifies the formula for the conjugate functional p' .

We now turn to the function q' . Consider first a point $(\xi, \eta) \notin \bar{\mathcal{T}}$. Since $\bar{\mathcal{T}}$ is a closed subspace, it follows from the separation theorem (strict form) that there exists $(\mu', \gamma') \in C^*(D \times \Omega) \times C^*(D)$ such that

$$(2.28) \quad \langle \mu', \xi \rangle + \langle \gamma', \eta \rangle < 0$$

and

$$(2.29) \quad \langle \mu', \phi_t(t, x) + \phi_x(t, x) \cdot v \rangle + \langle \gamma', -\phi \rangle = 0$$

for all $\phi \in C^1(\mathbb{R}^{1+n})$.

Now we know that the generalized flow $(\bar{\mu}, \bar{\gamma})$ corresponding to the admissible trajectory (\bar{t}_1, \bar{x}) satisfies

$$(2.30) \quad \langle \bar{\mu}, \phi_t(t, x) + \phi_x(t, x) \cdot v \rangle + \langle \bar{\gamma}, -\phi \rangle = -\phi(t_0, x_0)$$

for all $\phi \in C^1(\mathbb{R}^{1+n})$. Note that, by (2.29) and (2.30),

$$(\bar{\mu}, \bar{\gamma}) + \alpha(\mu', \gamma') \in \mathcal{T}$$

for every $\alpha \in \mathbb{R}$. It follows that

$$\begin{aligned} q'(\xi, \eta) &= \inf \{ \langle (\mu, \gamma), (\xi, \eta) \rangle : (\mu, \gamma) \in Q \} \\ &\leq \langle (\bar{\mu}, \bar{\gamma}), (\xi, \eta) \rangle + \inf_{\alpha \in \mathbb{R}} \langle (\mu', \gamma'), (\xi, \eta) \rangle \alpha \\ &= -\infty, \end{aligned}$$

by (2.28). We conclude that $q'(\xi, \eta) = -\infty$.

It remains to consider a point $(\xi, \gamma) \in \bar{\mathcal{T}}$. Since $\bar{\mathcal{T}}$ is the strong closure of \mathcal{T} , there exists a sequence of functions $\{\phi_i\}$ in $C^1(\mathbb{R}^{1+n})$ such that

$$(2.31) \quad \phi_i^i(t, x) + \phi_x^i(t, x) \cdot v \rightarrow \xi(t, x) \quad \text{and} \quad -\phi^i(t, x) \rightarrow \eta(t, x)$$

uniformly on $D \times \Omega$ and D , respectively.

For any $(\mu, \gamma) \in Q$, we have

$$\begin{aligned} \langle (\mu, \gamma), (\xi, \eta) \rangle &= \lim_i \langle (\mu, \gamma), (\phi_i^i(t, x) + \phi_x^i(t, x) \cdot v, -\phi^i(t, x)) \rangle \\ &= \lim_i -\phi^i(t_0, x_0) \\ &= \eta(t_0, x_0), \end{aligned}$$

by (2.31). It follows that

$$q'(\xi, \eta) = \inf \{ \langle (\mu, \gamma), (\xi, \eta) \rangle : (\mu, \gamma) \in Q \} = \eta(t_0, x_0).$$

q' therefore takes the form asserted in the lemma. \square

The formulae we have just derived for the conjugate functionals p' and q' lead to a simple characterization of the supremum cost for the dual problem.

LEMMA 2.9. *It holds that*

$$\sup \{ p'(\xi, \eta) - q'(\xi, \eta) : (\xi, \eta) \in C(D \times \Omega) \times C(D) \} = \sup \{ \phi(t_0, x_0) \},$$

where the supremum on the right is taken over elements $\phi \in C^1(\mathbb{R}^{1+n})$ that satisfy

$$\min_{v \in F(t, x)} \{ \phi_t(t, x) + \phi_x(t, x) \cdot v + l((t, x, v)) \} \geq 0 \quad \text{for } (t, x) \in A$$

and

$$\phi(t, x) \leq g(t, x) \quad \text{for all } (t, x) \in C.$$

Proof. By Lemma 2.8,

$$\sup \{ q'(\xi, \eta) - p'(\xi, \eta) : (\xi, \eta) \in C(D \times \Omega) \times C(D) \} = \sup \{ m(\xi, \eta) : (\xi, \eta) \in \bar{\mathcal{T}} \},$$

where m is the function

$$m(\xi, \eta) := \eta(t_0, x_0) - \max_{A \times \text{graph}\{F\}} (\xi - l)^+(T - t_0) - \max_C (\eta - g)^+,$$

and where $\bar{\mathcal{T}}$ is the closure of the set \mathcal{T} defined in Lemma 2.8. Since, however, m is a continuous function, the supremum on the right side is unaltered if we replace $\bar{\mathcal{T}}$ by \mathcal{T} . It follows that

$$\sup \{ q' - p' \} = \sup_{\phi \in C(\mathbb{R}^{1+n})} \tilde{m}(\phi),$$

where

$$\tilde{m}(\phi) := -\phi(t_0, x_0) - \max_{A \times \text{graph}\{F\}} (\phi_t + \phi_x \cdot v - l)^+(T - t_0) - \max_C (-\phi - g)^+.$$

For any $\phi \in C^1(\mathbb{R}^{1+n})$, define

$$c_1 := \max_{A \times \text{graph}\{F\}} (\phi_t + \phi_x \cdot v - l)^+ \quad \text{and} \quad c_2 := \max_C (-\phi - g)^+$$

and consider the C^1 function

$$\tilde{\phi}(t, x) := \phi(t, x) + c_2 - c_1(t - t_0).$$

We find that

$$\tilde{m}(\tilde{\phi}) \geq \tilde{m}(\phi),$$

$$\tilde{\phi}_t(t, x) + \tilde{\phi}_x(t, x) \cdot v - l(t, x, v) \leq 0 \quad \text{for all } (t, x, v) \in \text{graph}\{F\} \cap A,$$

and

$$-\tilde{\phi}(t, x) \leq g(t, x) \quad \text{for all } (t, x) \in C.$$

It follows that

$$\begin{aligned} & \sup \{ p' - q' \} \\ &= \sup \{ -\phi(t_0, x_0) : \phi \in C^1(\mathbb{R}^{1+n}), \phi_t + \phi_x \cdot v - l \leq 0 \\ & \quad \text{on } \text{graph}\{F\} \cap A, -\phi \leq g \text{ on } C \} \\ &= \sup \{ \phi(t_0, x_0) : \phi \in C^1(\mathbb{R}^{1+n}), \phi_t + \phi_x \cdot v + l \geq 0 \\ & \quad \text{on } \text{graph}\{F\} \cap A, \phi \leq g \text{ on } C \}. \quad \square \end{aligned}$$

We now return to the proof of Theorem 2.1. p' is a continuous function on $C(D \times \Omega) \times C(D)$, and q' is, of course, proper. By Proposition 2.7,

$$\inf \{W\} \quad (= \inf \{p - q\}) = \sup \{q' - p'\}.$$

Assembling this inequality with the assertions of Theorem 2.3 and Lemma 2.9 we see that (P) has a solution and

$$\inf \{P\} = \sup \{\phi(t_0, x_0)\},$$

where the supremum on the right is taken over functions $\phi \in C^1(\mathbb{R}^{1+n})$ satisfying the constraints listed in the statement of Theorem 2.1. This what we intended to prove.

REFERENCES

- [1] R. B. ASH, *Measure, Integration and Functional Analysis*, Academic Press, New York, 1972.
- [2] F. H. CLARKE AND R. B. VINTER, *Local optimality conditions and Lipschitzian solutions to the Hamilton-Jacobi equation*, SIAM J. Control Optim., 21 (1983), pp. 856-870.
- [3] W. H. FLEMING, *Generalized solutions and convex duality in optimal control*, in Partial Differential Equations and the Calculus of Variations: Essays in Honor of Ennio De Giorgi, Progress in Nonlinear Different Equations and their Applications, Birkhauser, Boston, 1989.
- [4] W. H. FLEMING AND D. VERMES, *Convex duality approach to the optimal control of diffusions*, SIAM J. Control Optim., 27 (1989), pp. 1136-1155.
- [5] A. D. IOFFE, *Convex problems occurring in variational problems and the absolute minimum problem*, Mat. Sb., 88 (1972), pp. 191-208.
- [6] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of Extremal Problems*, Studies in Mathematics and its Applications, 6, North-Holland, Amsterdam, 1979.
- [7] R. KLÖTZER, *On a general conception of duality in optimal control*, Lecture Notes Math., 703 (1979), pp. 189-196.
- [8] V. L. LEVIN AND A. A. MILYUTIN, *The problem of mass transfer with a discontinuous cost function, and a mass statement of the duality theorem for convex extremal problems*, Uspekhi Mat. Nauk, 34 (1979), pp. 3-68, Russian Math Surveys, 34 (1978), pp. 1-78.
- [9] P.-L. LIONS, *Generalized solutions of Hamilton Jacobi Equations*, Research Notes in Mathematics 69, Pitman, Boston, 1982.
- [10] R. T. ROCKAFELLAR, *An extension of Fenchel's duality theorem for convex problems*, Duke Math. J., 33 (1966), pp. 81-90.
- [11] R. B. VINTER, *Weakest conditions for existence of Lipschitz continuous Krotov functions in optimal control theory*, SIAM J. Control Optim., 21 (1983), pp. 235-245.
- [12] R. B. VINTER AND R. M. LEWIS, *The equivalence of strong and weak formulations for certain problems in optimal control*, SIAM J. Control Optim., 16 (1978), pp. 546-570.
- [13] ———, *A necessary and sufficient condition for optimality of dynamic programming type, making no a priori assumptions on the control*, SIAM J. Control Optim. 16 (1978), pp. 571-583.
- [14] R. B. VINTER AND P. WOLENSKI, *Hamilton Jacobi theory for problems with data measurable in time*, SIAM J. Control Optim., 28 (1990), pp. 1404-1419.
- [15] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, PA, 1969.
- [16] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

INFINITE-DIMENSIONAL VOLTERRA-STIELTJES EVOLUTION EQUATIONS AND RELATED OPTIMAL CONTROL PROBLEMS*

JIONGMIN YONG†

Abstract. Due to impulse control problems, a controlled Volterra–Stieltjes system in some Banach space is studied. In the system, the vector measure appearing in the Stieltjes-type integral (called Young integral here) is taken as a part of control action. In this paper, a general theory for infinite-dimensional Volterra–Stieltjes evolution equations is presented, and, for an optimal control problem governed by such type controlled system with a general endpoint constraint, a Pontryagin-type maximum principle is proved.

Key words. Young integral, Volterra–Stieltjes integral equation, distributed parameter system, measure control, maximum principle, optimal control

AMS(MOS) subject classifications. 49B27, 93C25, 93C22, 34G20

1. Introduction. Let us first give a motivation of the optimal control problem studied in this paper. Let X be a Banach space and $A: \mathcal{D}(A) \subset X \rightarrow X$ be the infinitesimal generator of some C_0 -semigroup e^{At} on X . We consider a controlled evolution system

$$(1.1) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + g(t, x(t), u(t)), & t \in [0, T], \\ x(0) &= x_0. \end{aligned}$$

The state of the system is usually understood as the mild solution of (1.1). Namely, the state $x(\cdot)$ of the system satisfies the following Volterra integral equation:

$$(1.2) \quad x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}g(s, x(s), u(s)) \, ds, \quad t \in [0, T].$$

Now suppose that we have another control action—an impulse control; i.e., at moment $t = \tau_i$, $i \geq 1$, we make an impulse ξ_i to the state $x(\tau_i - 0)$. We refer $\{(\tau_i, \xi_i) | i \geq 1\}$ as an “impulse control.” Thus we have that

$$x(\tau_i) = x(\tau_i - 0) + \xi_i, \quad i \geq 1.$$

Then the state $x(\cdot)$ formally satisfies the following evolution equation:

$$(1.3) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + g(t, x(t), u(t)) + \sum_{i \geq 1} \xi_i \delta(t - \tau_i), & t \in [0, T], \\ x(0) &= x_0, \end{aligned}$$

where $\delta(\cdot)$ is the δ -function. Similar to (1.2), we understand the state $x(\cdot)$ of system (1.3) as the solution of the following integral equation:

$$(1.4) \quad \begin{aligned} x(t) &= e^{At}x_0 + \int_0^t e^{A(t-s)}g(s, x(s), u(s)) \, ds \\ &+ \sum_{i \geq 1} e^{A(t-\tau_i)}\xi_i \chi_{[\tau_i, \infty)}(t), & t \in [0, T]. \end{aligned}$$

* Received by the editors January 24, 1990; accepted for publication (in revised form) September 12, 1991. This work was partially supported by Chinese National Science Foundation grants 0188416 and 1880414 and Chinese State Education Commission National Science Foundation grant 9024617.

† Department of Mathematics, Fudan University, Shanghai 200433, China.

It is more reasonable that, in general, the impulse ξ_i at time $t = \tau_i$ should also depend on the state $x(\tau_i - 0)$, which is the state of the system before making impulse ξ_i . Thus it is more natural to consider the following state equation (compare with (1.3)):

$$(1.5) \quad \begin{aligned} dx(t) &= (Ax(t) + g(t, x(t), u(t))) dt \\ &\quad + g_0(t, x(t-0), u(t)) d\xi(t), \quad t \in [0, T], \\ x(0) &= x_0, \end{aligned}$$

where $\xi(t) = \sum_{i \geq 1} \xi_i \chi_{[\tau_i, \infty)}(t)$, $t \in [0, \infty)$. If we let

$$G(t, x, u) = \begin{pmatrix} g(t, x, u) \\ g_0(t, x, u) \end{pmatrix}, \quad \eta(t) = \begin{pmatrix} t \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \xi(t) \end{pmatrix} = \begin{pmatrix} t \\ \xi(t) \end{pmatrix},$$

then (1.5) can be written as

$$(1.6) \quad \begin{aligned} dx(t) &= Ax(t) dt + G(t, x(t-0), u(t)) d\eta(t), \quad t \in [0, T], \\ x(0) &= x_0. \end{aligned}$$

From the above, we see that it is reasonable to consider the following controlled system (formally):

$$(1.7) \quad \begin{aligned} dx(t) &= Ax(t) dt + F(t, x(t-0), u(t)) d\mu(t), \quad t \in [0, T], \\ x(0) &= x_0, \end{aligned}$$

with some $\mathcal{L}(Z, X)$ -valued function F and some vector-valued measure $\mu(\cdot)$, which together with $u(\cdot)$ will be considered as control actions. We specify the meaning of (1.7) in § 2. Associated with (1.7), we are given a cost functional

$$(1.8) \quad J(x(\cdot), u(\cdot), \mu(\cdot)) = \int_0^T \langle f(t, x(t-0), u(t)), \mu(dt) \rangle.$$

Again, we give the precise meaning of the above later in the paper. Clearly, in the case where there is no impulse, (1.8) is reduced to the usual cost functional of Lagrange form. We should note, at this moment, that the functional (1.8) is adopted from [32]. It is not the most interesting case for the optimal impulse control problems since, in those problems, the cost functional is not necessarily linear in the impulse control variable $\mu(\cdot)$. For details, see [34] and [37] (see also [35] for the finite-dimensional case).

The control problem we study in this paper is minimizing functional (1.8) over some class of admissible controls, subject to the state equation (1.7) and an endpoint constraint for the state of the following type:

$$(1.9) \quad (x(0), x(T)) \in \Omega \subset X \times X.$$

In [31], [32], a similar problem in finite-dimensional spaces was studied. It is immediate that the first major difference between this paper and [31], [32] is the issue of whether the coefficient of $\mu(\cdot)$ depends on the state $x(\cdot)$. Second, we are in infinite-dimensional space, and we have a general endpoint constraint (1.9), which is different from the separated endpoint constraint case (see [23] for comments). On the other hand, we should note that the operator-valued function $F(t, x, u)$ must be assumed Fréchet differentiable in x , which is slightly more restrictive than [32] for the finite-dimensional case (in [32] only the Lipschitz continuity in x was assumed). The removal of this restriction, as well as the study of more general cost functionals, is expected in our future work. For some special cases, the idea of [36] might be useful (see also [4]). We refer to [12] for some relevant results.

Impulse control problems have frequently been discussed. We refer to [6], [7], [25] for finite-dimensional stochastic systems, to [5], [35] for finite-dimensional deterministic systems, and to [34], [37] for infinite-dimensional cases. We refer to [8]–[10], [30], [33] for some classical optimal control problems and to [1]–[4], [17], [22]–[24] for distributed parameter system cases.

The so-called Young integral for scalar functions was introduced by L. C. Young in 1914 [38]. We refer to [18], [19] for the theory of Volterra–Stieltjes integral equations involving the Young integral (for the scalar-valued case). Some relevant results concerning the Volterra–Stieltjes integral equations (involving some other types of integrals) can be found in [21], [26], [28]. We should note that the integrals involved in [31], [32] were of the Bochner (i.e., the Lebesgue) type. We adopt the Young integral because such an integral inherits the exact “discontinuity” of the measure with respect to which the integral is taken. This is the case for the impulse control problems. It seems that a similar property of the Bochner-type integral is not as good as the one of the Young integral (see Remark 3.1).

The main results of this paper consist of the study of the infinite-dimensional Volterra–Stieltjes integral equations and the Pontryagin-type maximum principle for the related optimal control problems. We organize the paper as follows. In § 2 we present a theory for the infinite-dimensional Volterra–Stieltjes integral equations. In § 3 we state the optimal control problem and the maximum principle. We give some auxiliary results in § 4. Section 5 is devoted to proving the maximum principle. Some relevant results concerning the vector measures and Young integrals are collected in the Appendix.

2. Volterra–Stieltjes integral equations. In this section, we present some basic results concerning the Volterra–Stieltjes equations involving the Young integral. We refer to the Appendix for relevant results concerning the Young integral. We let X and Z be Banach spaces, K be a subset of Z , and

$$\begin{aligned} \mu(\cdot) &\in BV_0([0, T]; K) \\ &\equiv \{v(\cdot) : [0, T] \rightarrow K \mid v(\cdot) \text{ is of bounded variation and } v(0) = 0\}. \end{aligned}$$

We denote the bounded variational vector measure associated with μ by $\bar{\mu}$. The total variations of μ and $\bar{\mu}$ are denoted by $|\mu|$ and $|\bar{\mu}|$, respectively. Let

$$\Delta = \{(t, \tau) \in [0, T] \times [0, T] \mid 0 \leq \tau < t \leq T\}$$

and let $\bar{\Delta}$ be the closure of Δ . We consider the following integral equation:

$$(2.1) \quad x(t) = \varphi(t) + \int_0^t G(t, \tau, x(\tau-0)) d\mu(\tau), \quad t \in [0, T].$$

Let us assume that the following statements are true.

Assumption 1. The function $\varphi(\cdot) \in C_\mu([0, T]; X)$ holds (see the Appendix).

Assumption 2. The function $G : \bar{\Delta} \times X \rightarrow \mathcal{L}(Z, X)$ satisfies the following conditions:

(i) There exists a constant L such that

$$(2.2) \quad \|G(t, \tau, x)\|_{\mathcal{L}(Z, X)} \leq L(1 + |x|) \quad \forall (t, \tau) \in \bar{\Delta}, \quad x \in X,$$

$$(2.3) \quad \|G(t, \tau, x) - G(t, \tau, \hat{x})\|_{\mathcal{L}(Z, X)} \leq L|x - \hat{x}| \quad \forall (t, \tau) \in \bar{\Delta}, \quad x, \hat{x} \in X;$$

(ii) For any $\tau \in [0, T]$ and $x \in X$, the map $G(\cdot, \tau, x)$ is continuous on $[\tau, T]$.

We have the following result.

THEOREM 2.1. Let $\mu(\cdot) \in BV_0([0, T]; K)$, and let Assumptions 1 and 2 hold. Then (2.1) has a unique solution $x(\cdot) \in C_\mu([0, T]; X)$.

To prove this theorem, we need the following lemma, the proof of which can be found in [18].

LEMMA 2.2. *Let $\lambda(\cdot) \in BV_0([0, T]; \mathbb{R})$ and $h_0 \in \mathbb{R}$. Then there exists a unique solution $h(\cdot) \in BV([0, T]; \mathbb{R})$ of the following integral equation:*

$$(2.4) \quad h(t) = h_0 + \int_0^t h(\tau - 0) d\lambda(\tau), \quad t \in [0, T].$$

Moreover, the solution $h(\cdot)$ can be written as

$$(2.5) \quad h(t) = h_0 e^{\lambda(t)} [1 + \lambda(t) - \lambda(t-0)] e^{-[\lambda(t) - \lambda(t-0)]} \cdot \prod_{0 \leq \tau < t} [1 + \lambda(\tau + 0) - \lambda(\tau - 0)] e^{-[\lambda(\tau + 0) - \lambda(\tau - 0)]}, \quad t \in [0, T].$$

COROLLARY 2.3. *Let $\lambda(\cdot) \in BV_0([0, T]; \mathbb{R})$ be increasing and $h_0 \geq 0$. Then the unique solution $h(\cdot)$ of (2.4) satisfies*

$$(2.6) \quad h_0 \leq h(t) \leq h_0 e^{\lambda(t)}, \quad t \in [0, T].$$

Proof. By (2.5), we see that (since $h_0 \geq 0$ and $\lambda(\cdot)$ is increasing)

$$(2.7) \quad h(t) \geq 0 \quad \forall t \in [0, T].$$

Thus by (2.4) we see that

$$(2.8) \quad h(t) \geq h_0 \quad \forall t \in [0, T].$$

On the other hand, since $(1+a)e^{-a} \leq 1$ for all $a \geq 0$, we obtain the other half of (2.6). \square

Proof of Theorem 2.1. First, by Lemma 2.2 we let $h(\cdot)$ be the unique solution of the following equation:

$$(2.9) \quad h(t) = 1 + 2L \int_0^t h(\tau - 0) d|\mu|(\tau), \quad t \in [0, T].$$

Then by Corollary 2.3 we know that

$$(2.10) \quad 1 \leq h(t) \leq e^{2L|\mu|(t)}, \quad t \in [0, T].$$

Next, we define

$$(2.11) \quad \|x(\cdot)\|_h = \sup_{t \in [0, T]} h(t)^{-1} |x(t)|.$$

By definition, we see that $C_\mu([0, T]; X)$ is complete under this norm. Then, for any $x(\cdot) \in C_\mu([0, T]; X)$, we define

$$(2.12) \quad (\mathcal{S}x)(t) = \varphi(t) + \int_0^t G(t, \tau, x(\tau - 0)) d\mu(\tau), \quad t \in [0, T].$$

Since $\varphi(\cdot) \in C_\mu([0, T]; X)$, by Proposition A.8 of the Appendix, we see that

$$(2.13) \quad \mathcal{S}: C_\mu([0, T]; X) \rightarrow C_\mu([0, T]; X).$$

On the other hand, for any $x(\cdot), \hat{x}(\cdot) \in C_\mu([0, T]; X)$, we have that

$$\begin{aligned} |(\mathcal{S}x)(t) - (\mathcal{S}\hat{x})(t)| &\leq L \int_0^t |x(\tau - 0) - \hat{x}(\tau - 0)| d|\mu|(\tau) \\ &\leq L \|x(\cdot) - \hat{x}(\cdot)\|_h \int_0^t h(\tau - 0) d|\mu|(\tau) \\ &= \frac{h(t) - 1}{2} \|x(\cdot) - \hat{x}(\cdot)\|_h \quad \forall t \in [0, T]. \end{aligned}$$

Hence $\|(\mathcal{S}x)(\cdot) - (\mathcal{S}\hat{x})(\cdot)\|_h \leq \frac{1}{2} \|x(\cdot) - \hat{x}(\cdot)\|_h$. Then, by the Banach fixed point theorem, there exists a unique solution $x(\cdot) \in C_\mu([0, T]; X)$ of (2.1). \square

The next results give some basic properties of the solution of (2.1). The proofs are simple.

PROPOSITION 2.4. *Let Assumption 2 hold and let $\varphi(\cdot), \hat{\varphi}(\cdot) \in C_\mu([0, T]; X)$. Let $x(\cdot)$ and $\hat{x}(\cdot)$ be the solutions of (2.1) corresponding to $\varphi(\cdot)$ and $\hat{\varphi}(\cdot)$, respectively. Then, for any $\beta > 1$,*

$$(2.14) \quad |x(t) - \hat{x}(t)| \leq \frac{\beta}{\beta - 1} \left[\sup_{0 \leq \tau \leq t} |\varphi(\tau) - \hat{\varphi}(\tau)| \right] e^{\beta L|\mu|(t)}, \quad t \in [0, T],$$

$$(2.15) \quad |x(t)| \leq \frac{\beta}{\beta - 1} \left[\sup_{0 \leq \tau \leq t} |\varphi(\tau)| + L|\mu|(t) \right] e^{\beta L|\mu|(t)}, \quad t \in [0, T].$$

Assumption 3. There exist functions $\omega_G: [0, T] \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and $\theta: [0, T] \rightarrow \mathbb{R}^+$ with the properties that $\omega_G(0, r) = 0$, for all $r \in \mathbb{R}^+$, $\omega_G(t, \cdot)$ and $\omega_G(\cdot, r)$ are nondecreasing and $\theta(\cdot)$ is $|\bar{\mu}|$ -integrable, such that

$$(2.16) \quad \|G(t, \tau, x) - G(s, \tau, x)\|_{\mathcal{L}(Z, X)} \leq \omega_G(\rho_\mu(t, s), |x|) \theta(\tau) \\ \forall \tau \in [0, T], \quad t, s \in [\tau, T], \quad x \in X.$$

PROPOSITION 2.5. *Let Assumptions 1–3 hold. Then, for the unique solution $x(\cdot)$ of (2.1), we have that*

$$(2.17) \quad |x(t) - x(s)| \leq |\varphi(t) - \varphi(s)| + L \left(1 + \sup_{0 \leq \tau \leq t} |x(\tau)| \right) (|\mu|(t) - |\mu|(s)) \\ + \omega_G \left(\rho_\mu(t, s), \sup_{0 \leq \tau \leq t} |x(\tau)| \right) \int_0^s \theta(\tau) d|\mu|(\tau) \\ \forall 0 \leq s \leq t \leq T.$$

In particular, if $\varphi(\cdot) \in BV([0, T]; X)$ and

$$(2.18) \quad \omega_G(t, r) = t\bar{\omega}(r) \quad \forall (t, r) \in [0, T] \times \mathbb{R}^+,$$

for some nondecreasing function $\bar{\omega}: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ with $\bar{\omega}(0) = 0$, then $x(\cdot) \in BV([0, T]; X)$.

We should note that, in the case we study, the function $\varphi(\cdot)$ is not necessarily bounded variational and that (2.18) does not hold (in general), either. Thus we are not expected to obtain bounded variational solutions (although they are μ -continuous). However, in the finite-dimensional case, we do have $x(\cdot) \in BV([0, T]; X)$, provided that (2.16) and (2.18) hold, which is not too restrictive. Also, we should note that Assumption 1 is just for simplicity. It can be replaced by

$$\varphi(\cdot) \in D([0, T]; X) = \{ \varphi: [0, T] \rightarrow X \mid \varphi(t+0) \text{ and } \varphi(t-0) \text{ exist} \\ \text{for all } t \in [0, T] \},$$

and we seek the solutions of (2.1) in $D([0, T]; X)$ instead. The results will be similar.

The remainder of this section is devoted to the study of the following linear equation:

$$(2.19) \quad x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)} B(\tau, x(\tau-0), d\mu(\tau)) \\ + \int_0^t e^{A(t-\tau)} \bar{G}(\tau) d\nu(\tau), \quad t \in [0, T].$$

We make the following assumptions.

Assumption 4. The operator $A: \mathcal{D}(A) \subset X \rightarrow X$ generates a C_0 -semigroup e^{At} on X and, for some constants $M \geq 1$ and $\omega \in \mathbb{R}$,

$$(2.20) \quad \|e^{At}\|_{\mathcal{L}(X)} \leq M e^{\omega t}, \quad t \geq 0.$$

Assumption 5. Function $B: [0, T] \rightarrow \tilde{\mathcal{B}}(X \times Z; X)$ is uniformly Borel-measurable, with $\tilde{\mathcal{B}}(X \times Z; X)$ being the set of all bounded bilinear forms from $X \times Z$ to X (see the Appendix), and $\|B(\tau)\|_{\tilde{\mathcal{B}}}$ is bounded by L_0 .

Assumption 6. $\nu(\cdot) \in BV_0([0, T]; Y)$ and $\bar{G}(\cdot) \in B_\nu([0, T]; \mathcal{L}(Y, X))$, with Y being some Banach space.

Under Assumptions 4 and 5, we know that there exists a unique solution $x(\cdot)$ of (2.19). Our next goal is to give a variation of constants formula. To this end, let us define the evolution operators $\Phi, \Theta: \bar{\Delta} \rightarrow \mathcal{L}(X)$ as follows:

$$(2.21) \quad \begin{aligned} \Phi(t, s)x_0 &= e^{A(t-s)}x_0 + \int_s^t e^{A(t-\tau)}B(\tau, \Phi(\tau-0, s)x_0, d\mu(\tau)), \\ 0 \leq s \leq t \leq T, \quad x_0 \in X, \end{aligned}$$

$$\Phi(s-0, s) = I, \quad 0 \leq s \leq T;$$

$$(2.22) \quad \begin{aligned} \Theta(t, s)x_0 &= \int_s^t e^{A(t-\tau)}B(\tau, \Theta(\tau-0, s)x_0, d\mu(\tau)), \\ 0 \leq s \leq t \leq T, \quad x_0 \in X, \end{aligned}$$

$$\Theta(s-0, s) = -I, \quad 0 \leq s \leq T.$$

It is clear that the above operator-valued functions are well defined. Also, we see that

$$(2.23) \quad \Theta(s, s) = 0, \quad \Theta(s+0, s) = B(s, -I, \mu(s+0) - \mu(s)).$$

Thus, if $\mu(\cdot)$ is right-continuous at some point $s \in [0, T]$, then

$$(2.24) \quad \Theta(t, s) = 0 \quad \forall 0 \leq s \leq t \leq T.$$

THEOREM 2.6 (variation of constants formula). *The unique solution $x(\cdot)$ of (2.19) can be represented by*

$$(2.25) \quad x(t) = \Phi(t, 0)x_0 + \int_0^t (\Phi(t, s) + \Theta(t, s))\bar{G}(s) d\nu(s), \quad t \in [0, T].$$

Proof. For simplicity, we denote

$$(2.26) \quad \Psi(t, s) = \Phi(t, s) + \Theta(t, s), \quad 0 \leq s \leq t \leq T.$$

Then we have that

$$(2.27) \quad \begin{aligned} \Psi(t, s)x_0 &= e^{A(t-s)}x_0 + \int_s^t e^{A(t-\tau)}B(\tau, \Psi(\tau-0, s)x_0, d\mu(\tau)), \\ 0 \leq s \leq t \leq T, \quad x_0 \in X, \\ \Psi(s-0, s) &= 0, \quad 0 \leq s \leq T. \end{aligned}$$

Now we set

$$(2.28) \quad y(t) = \Phi(t, 0)x_0 + \int_0^t \Psi(t, s) \bar{G}(s) d\nu(s), \quad t \in [0, T].$$

Then we have that

$$(2.29) \quad y(\tau - 0) = \Phi(\tau - 0, 0)x_0 + \int_{[0, \tau)} \Psi(\tau - 0, s) \bar{G}(s) \bar{\nu}(ds), \quad \tau \in [0, T].$$

Let us observe the following:

$$\begin{aligned}
 & \int_0^t e^{A(t-\tau)} B(\tau, y(\tau - 0), d\mu(\tau)) \\
 &= \int_0^t e^{A(t-\tau)} B(\tau, \Phi(\tau - 0, 0)x_0, d\mu(\tau)) \\
 & \quad + \int_0^t e^{A(t-\tau)} B\left(\tau, \int_{[0, \tau)} \Psi(\tau - 0, s) \bar{G}(s) \bar{\nu}(ds), d\mu(\tau)\right) \\
 &= \Phi(t, 0)x_0 - e^{At}x_0 + \int_{[0, t)} e^{A(t-\tau)} B\left(\tau, \int_{[0, \tau)} \Psi(\tau - 0, s) \bar{G}(s) \bar{\nu}(ds), \bar{\mu}(d\tau)\right) \\
 & \quad + B\left(t, \int_{[0, t)} \Psi(t - 0, s) \bar{G}(s) \bar{\nu}(ds), \mu(t) - \mu(t - 0)\right) \\
 (2.30) \quad &= \Phi(t, 0)x_0 - e^{At}x_0 + \int_{[0, t)} \int_{(s, t)} e^{A(t-\tau)} B(\tau, \Psi(\tau - 0, s) \bar{G}(s) \bar{\nu}(ds), \bar{\mu}(d\tau)) \\
 & \quad + \int_{[0, t)} B(t, \Psi(t - 0, s) \bar{G}(s) \bar{\nu}(ds), \mu(t) - \mu(t - 0)) \\
 &= \Phi(t, 0)x_0 - e^{At}x_0 + \int_0^t \int_s^t e^{A(t-\tau)} B(\tau, \Psi(\tau - 0, s) \bar{G}(s) d\nu(s), d\mu(\tau)) \\
 &= \Phi(t, 0)x_0 - e^{At}x_0 + \int_0^t (\Psi(t, s) - e^{A(t-s)}) \bar{G}(s) d\nu(s) \\
 &= y(t) - e^{At}x_0 - \int_0^t e^{A(t-s)} \bar{G}(s) d\nu(s).
 \end{aligned}$$

Hence (2.25) follows from the uniqueness of the solutions of (2.19). \square

In the above, the crucial step is to use the Fubini-type theorem (Theorem A.9) to exchange the integral order. For the case where $\mu(\cdot) \in \mathcal{M}([0, T]; Z)$ and $\nu(\cdot) \in \mathcal{M}([0, T]; Y)$, the proof looks clearer.

LEMMA 2.7. *The operator-valued function $\Psi(\cdot, \cdot)$ also satisfies the following:*

$$(2.31) \quad \Psi(t, s)x_0 = e^{A(t-s)}x_0 + \int_s^t \Psi(t, \tau) B(\tau, e^{A(\tau-s)}x_0, d\mu(\tau)),$$

$$0 \leq s \leq t \leq T.$$

Proof. We let the right-hand side of (2.31) be $\Psi(t, s)$. Then, by noting Theorem A.9 and the fact that $\Psi(s-0, s) = 0$, we have that

$$\begin{aligned}
 & \int_s^t \Psi(t, \tau) B(\tau, e^{A(\tau-s)} x_0, d\mu(\tau)) \\
 &= \int_s^t e^{A(t-\tau)} B(\tau, e^{A(\tau-s)} x_0, d\mu(\tau)) \\
 & \quad + \int_s^t \int_\tau^t e^{A(t-\tau)} B(r, \Psi(r-0, \tau) B(\tau, e^{A(\tau-s)} x_0, d\mu(\tau)), d\mu(r)) \\
 &= \int_s^t e^{A(t-\tau)} B(\tau, e^{A(\tau-s)} x_0, d\mu(\tau)) \\
 (2.32) \quad & \quad + \int_s^t \int_s^\tau e^{A(t-\tau)} B(r, \Psi(r-0, \tau) B(\tau, e^{A(\tau-s)} x_0, d\mu(\tau)), d\mu(r)) \\
 &= \int_s^t e^{A(t-\tau)} B(\tau, e^{A(\tau-s)} x_0 \\
 & \quad + \int_s^\tau \Psi(\tau-0, r) B(r, e^{A(\tau-s)} x_0, d\mu(r)), d\mu(\tau)) \\
 &= \int_s^t e^{A(t-\tau)} B(\tau, \bar{\Psi}(\tau-0, s) x_0, d\mu(\tau)).
 \end{aligned}$$

By uniqueness, we obtain (2.31). \square

This result is very similar to that given in [11] for the case where $\mu(t) = t$ (see also [22]). This result will be used in deriving the adjoint system along the optimal pair in § 5.

3. Optimal control problem. In this section, based in the preliminary results of the previous section, we state our optimal control problem and the main results concerning the control problem of this paper. We let X and Z be two Banach spaces, K be a convex and closed cone in Z , and U be a metric space. We define

$$\mathcal{U} = \{u(\cdot) : [0, T] \rightarrow U \mid u(\cdot) \text{ is Borel-measurable}\}$$

and

$$\mathcal{M} = \{\mu_0(\cdot)\} + \mathcal{M}_0([0, T]; K),$$

where $\mathcal{M}_0([0, T]; K)$ is defined as in the Appendix, and $\mu_0(\cdot) \in BV_0([0, T]; Z)$. By Proposition A.1, we know that \mathcal{M} is convex. Next, let us make the following hypotheses.

Hypothesis 1. The dual X^* of X is strictly convex.

Hypothesis 2. This is identical to Assumption 4 (see § 2).

Hypothesis 3. The map $F : [0, T] \times X \times U \rightarrow \mathcal{L}(Z, X)$ satisfies the following conditions:

(i) For any $(x, u) \in X \times U$, the map $F(\cdot, x, u) : [0, T] \rightarrow \mathcal{L}(Z, X)$ is uniformly Borel-measurable. For any $(t, x) \in [0, T] \times U$, the map $F(t, \cdot, u) : X \rightarrow \mathcal{L}(Z, X)$ is continuously Fréchet differentiable. For any $(t, x) \in [0, T] \times X$, the map $F(t, x, \cdot) : U \rightarrow \mathcal{L}(Z, X)$ is continuous;

(ii) There exists a constant $L > 0$ such that

$$(3.1) \quad \|F(t, x, u)\|_{\mathcal{L}(Z, X)} \leq L(1 + |x|), \quad (t, x, u) \in [0, T] \times X \times U,$$

$$(3.2) \quad \|F(t, x, u) - F(t, \hat{x}, u)\|_{\mathcal{L}(Z, X)} \leq L|x - \hat{x}|, \quad (t, u) \in [0, T] \times U, \quad x, \hat{x} \in X.$$

Hypothesis 4. The map $f: [0, T] \times X \times U \rightarrow Z^*$ satisfies the following conditions. For any $(x, u) \in X \times U$, the map $f(\cdot, x, u): [0, T] \rightarrow Z^*$ is uniformly Borel-measurable. For any $(t, x) \in [0, T] \times X$, the map $f(t, \cdot, u): X \rightarrow Z^*$ is continuously Fréchet differentiable. For any $(t, x) \in [0, T] \times X$, the map $f(t, x, \cdot): U \rightarrow Z^*$ is continuous.

Hypothesis 5. The set Ω is convex and closed in $X \times X$.

Now, for any $x_0 \in X$ and a pair $(u(\cdot), \mu(\cdot)) \in \mathcal{U} \times \mathcal{M}$, the response of the controlled system is defined to be the unique solution of the following integral equation:

$$(3.3) \quad x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)} F(\tau, x(\tau-0), u(\tau)) d\mu(\tau), \quad t \in [0, T].$$

From the theory we presented in § 2, we see that the above equation makes sense. Moreover, by Theorem 3.1, we know that, under Hypotheses 2 and 3, there exists a unique solution $x(\cdot) \in C_\mu([0, T]; X)$ of (3.3) corresponding to the triplet $(x_0, u(\cdot), \mu(\cdot))$. System (3.3) is a mild form of (1.7). Thus we sometimes refer to the solution $x(\cdot)$ of (3.3) as the mild solution (1.7). We call $x(t)$ the state of our system at time t ; $x(\cdot)$ the trajectory of the system, and x_0 , $u(\cdot)$, and $\mu(\cdot)$ the initial state, continuous control, and measure control, respectively. We see that the measure control is an extended notion of the so-called impulse control [6], [7], [34], [35]. Hypothesis 1 is technical for later purposes. For the case where X is reflexive or separable, we can always change the norm of X to another equivalent one so that Hypothesis 1 holds (see [23] for comments).

Remark 3.1. It is very important to note that, by using the Young-type integral in (3.3), the trajectory of the system inherits the discontinuity of the measure control $\mu(\cdot)$. In fact, if we let $x(\cdot)$ be the solution of (3.3) corresponding to $(x_0, u(\cdot), \mu(\cdot))$, then, by Proposition A.8, we have that

$$x(t) = x(t-0) + F(t, x(t-0), u(t))[\mu(t) - \mu(t-0)], \quad t \in (0, T]$$

and

$$x(t+0) = x(t) + F(t, x(t-0), u(t))[\mu(t+0) - \mu(t)], \quad t \in (0, T].$$

The Bochner-type integral does not have such a property! As in [32], the Bochner- (or Lebesgue-) type integral was used, and, as a result, the trajectories had to be restricted to right-continuous functions. In our case, the trajectories are only of elements in the space $D([0, T]; X)$.

The payoff of our control problem is the following (which is the same as (1.8)):

$$(3.4) \quad J(x(\cdot), u(\cdot), \mu(\cdot)) = \int_0^T \langle f(t, x(t-0), u(t)), d\mu(t) \rangle.$$

The meaning of the right-hand side of (3.4) is clear if we regard the map f as a $\mathcal{L}(Z, \mathbb{R})$ -valued function. Now we are ready to state our optimal control problem.

PROBLEM C. We minimize functional (3.4) over all $(u(\cdot), \mu(\cdot)) \in \mathcal{U} \times \mathcal{M}$, subject to system (3.3) and the endpoint constraint

$$(3.5) \quad (x(0), x(T)) \in \Omega.$$

Next, let $(x^\#(\cdot), u^\#(\cdot), \mu^\#(\cdot))$ be an optimal solution to Problem C. We define

$$(3.6) \quad \mathcal{U}^\# = \{u(\cdot) \in \mathcal{U} \mid u(\tau) = u^\#(\tau), \text{ if } \mu^\#(\cdot) \text{ jumps at } \tau\}.$$

Note that $F: [0, T] \times X \times U \rightarrow \mathcal{L}(Z, X)$. Thus the Fréchet derivative F_x of F is a map

from $[0, T] \times X \times U$ to $\tilde{B}(Z \times X; X)$. We let

$$(3.7) \quad \begin{aligned} B(\tau, \xi, z) &= F_x(\tau, x^\#(\tau-0), u^\#(\tau); \xi)z \\ &\equiv \lim_{\varepsilon \rightarrow 0} \frac{F(\tau, x^\#(\tau-0) + \varepsilon \xi, u^\#(\tau))z - F(\tau, x^\#(\tau-0), u^\#(\tau))z}{\varepsilon} \\ \forall(\tau, \xi, z) &\in [0, T] \times X \times Z. \end{aligned}$$

Then we may define $B(\cdot, \cdot)^*: [0, T] \times Z \rightarrow \mathcal{L}(X^*)$ as follows:

$$(3.8) \quad \langle B(\tau, z)^* x^*, \xi \rangle = \langle B(\tau, \xi, z), x^* \rangle \quad \forall(\tau, \xi, z, x^*) \in [0, T] \times X \times Z \times X^*.$$

As in § 2, we let $\Phi(\cdot, \cdot)$, $\Theta(\cdot, \cdot)$, and $\Psi(\cdot, \cdot)$ be the evolution operators associated with the operator A and the bilinear form-valued function $B(\tau, \xi, z)$ (see (2.21), (2.22), and (2.26)). Then we define

$$(3.9) \quad \begin{aligned} \mathcal{R} &= \left\{ \int_0^T \Psi(T, \tau) [F(\tau, x^*(\tau-0), u(\tau)) - F(\tau, x^\#(\tau-0), u^\#(\tau))] d\mu^\#(\tau) \right. \\ &\quad \left. + \int_0^T \Psi(T, \tau) F(\tau, x^\#(\tau-0), u^\#(\tau)) d(\mu(\tau) - \mu^\#(\tau)) \right. \\ &\quad \left. \left| (u(\cdot), \mu(\cdot)) \in \mathcal{U}^\# \times \mathcal{M} \right\} \end{aligned}$$

and

$$(3.10) \quad \mathcal{Q} = \{y_1 - \Phi(T, 0)y_0 \mid (y_0, y_1) \in \Omega\}.$$

We introduce the Hamiltonian

$$(3.11) \quad \begin{aligned} H(\tau, x, u, \psi^0, \psi) &= \psi^0 f(\tau, x, u) + F(\tau, x, u)^* \psi \\ \forall(\tau, x, u, \psi^0, \psi) &\in [0, T] \times X \times \mathbb{R} \times X^*. \end{aligned}$$

We note that the function $H(\tau, x, u, \psi^0, \psi)$ is Z^* -valued.

Now we are ready to state our main result for Problem C.

THEOREM 3.2 (maximum principle). *Let Hypotheses 1-5 hold. Let $(x^\#(\cdot), u^\#(\cdot), \mu^\#(\cdot))$ be an optimal solution to Problem C. Let the set*

$$\mathcal{R} - \mathcal{Q} = \{r - q \mid r \in \mathcal{R}, q \in \mathcal{Q}\}$$

be finite-codimensional in X [17], [22], [23]. Then there exists a pair $(\psi(\cdot), \psi^0) \in D([0, T]; X^) \times \mathbb{R}$, such that*

$$(3.12) \quad \begin{aligned} \psi(t) &= e^{A^*(T-t)} \psi(T) + \int_t^T e^{A^*(\tau-t)} B(\tau, d\mu^\#(\tau))^* \psi(\tau) \\ &\quad + \psi^0 \int_t^T e^{A^*(\tau-t)} f_x(\tau, x^\#(\tau-0), u^\#(\tau))^* d\mu^\#(\tau) \quad \forall t \in [0, T], \end{aligned}$$

$$(3.13) \quad \psi^0 \leq 0,$$

$$(3.14) \quad (\psi(\cdot), \psi^0) \neq 0,$$

$$(3.15) \quad \int_0^T \langle H(\tau, x^\#(\tau-0), u^\#(\tau), \psi^0, \psi(\tau)), d\mu^\#(\tau) \rangle$$

$$= \max_{\mu(\cdot) \in \mathcal{M}} \int_0^T \langle H(\tau, x^\#(\tau-0), u^\#(\tau), \psi^0, \psi(\tau)), d\mu(\tau) \rangle,$$

$$(3.16) \quad \int_0^T \langle H(\tau, x^\#(\tau-0), u^\#(\tau), \psi^0, \psi(\tau)), d\mu^\#(\tau) \rangle$$

$$= \max_{u(\cdot) \in \mathcal{U}^\#} \int_0^T \langle H(\tau, x^\#(\tau-0), u(\tau), \psi^0, \psi(\tau)), d\mu^\#(\tau) \rangle,$$

$$(3.17) \quad \langle \psi(0), y_0 - x^\#(0) \rangle - \langle \psi(T), y_1 - x^\#(T) \rangle$$

$$- \langle \psi(T), \Theta(T, 0)(y_0 - x^\#(0)) \rangle \leq 0, \quad (y_0, y_1) \in \Omega.$$

Remark 3.3. Condition (3.17) is the transversality condition for the optimal solution of our problem. The appearance of the term $\langle \psi(T), \Theta(T, 0)(y_0 - x^\#(0)) \rangle$ is unexpected. This is caused by a possible jump of the measure control $\mu(\cdot)$ at $t=0$. It is interesting to note that in the case where

$$(3.18) \quad \mu^\#(0+0) = \mu^\#(0)$$

(i.e., where no jump of $\mu^\#(\cdot)$ appeared at $t=0$), from (2.24), we see that

$$(3.19) \quad \Theta(t, 0) = 0 \quad \forall t \in [0, T].$$

Thus condition (3.17) is reduced to the familiar form

$$(3.20) \quad \langle \psi(0), y_0 - x^\#(0) \rangle - \langle \psi(T), y_1 - x^\#(T) \rangle \leq 0, \quad (y_0, y_1) \in \Omega.$$

4. Some lemmas. In this section, we present some lemmas which are necessary for the proof of the maximum principle in the next section.

LEMMA 4.1. *Let Assumptions 4 and 5 hold. Let $x(\cdot)$ and $\hat{x}(\cdot)$ be the unique solutions of (3.3) corresponding to $(x_0, u(\cdot), \mu(\cdot))$ and $(\hat{x}_0, \hat{u}(\cdot), \hat{\mu}(\cdot))$, respectively. Then there exists a constant C , depending on the bounds of $|x_0|$, $|\hat{x}_0|$, $|\mu|(T)$, $|\hat{\mu}|(T)$ and L , such that*

$$(4.1) \quad |x(t) - \hat{x}(t)| \leq C[|x_0 - \hat{x}_0| + |\mu - \hat{\mu}|(t)$$

$$+ |\bar{\mu}|(\{u \neq \hat{u}\}) \wedge |\bar{\mu}|(\{u \neq \hat{u}\})], \quad t \in [0, T].$$

Proof. First, from (3.3), we see that

$$(4.2) \quad |x(t)| \leq M e^{\omega t} |x_0| + ML \int_0^t e^{\omega(t-\tau)} (1 + |x(\tau-0)|) d|\mu|(\tau).$$

Thus

$$(4.3) \quad e^{-\omega t} |x(t)| \leq M |x_0| + ML \int_0^t e^{-\omega\tau} d|\mu|(\tau)$$

$$+ ML \int_0^t e^{-\omega\tau} |x(\tau-0)| d|\mu|(\tau).$$

Set $\sigma(t) = e^{-\omega t} |x(t)|$. Next, let $\beta > 1$ and $h_\beta(\cdot)$ be the unique solution of

$$(4.4) \quad h_\beta(t) = 1 + \beta ML \int_0^t h_\beta(\tau-0) d|\mu|(\tau), \quad t \in [0, T].$$

Then

$$\begin{aligned}
 \sigma(t) &\leq M|x_0| + ML \int_0^t e^{-\omega\tau} d|\mu|(\tau) + ML \int_0^t \sigma(\tau-0) d|\mu|(\tau) \\
 (4.5) \quad &\leq M|x_0| + ML \int_0^t e^{-\omega\tau} d|\mu|(\tau) + \sup_{0 \leq \tau \leq t} [\sigma(\tau)h_\beta(\tau)^{-1}] \frac{h_\beta(t)-1}{\beta}.
 \end{aligned}$$

By Corollary 2.3, we have that

$$\begin{aligned}
 \sup_{0 \leq \tau \leq t} [\sigma(\tau)h_\beta(\tau)^{-1}] &\leq M|x_0| + ML(1 \vee e^{-\omega t})|\mu|(t) \\
 (4.6) \quad &+ \frac{1}{\beta} \sup_{0 \leq \tau \leq t} [\sigma(\tau)h_\beta(\tau)^{-1}].
 \end{aligned}$$

Hence, again by Proposition 2.5, we obtain

$$(4.7) \quad |x(t)| \leq \frac{\beta}{\beta-1} (M|x_0| + ML(1 \vee e^{-\omega t})|\mu|(t)) e^{\omega t + \beta ML|\mu|(t)}, \quad t \in [0, T].$$

Similarly,

$$(4.8) \quad |\hat{x}(t)| \leq \frac{\beta}{\beta-1} (M|\hat{x}_0| + ML(1 \vee e^{-\omega t})|\hat{\mu}|(t)) e^{\omega t + \beta ML|\hat{\mu}|(t)}, \quad t \in [0, T].$$

Then we have that

$$\begin{aligned}
 |x(t) - \hat{x}(t)| &\leq M e^{\omega t} |x_0 - \hat{x}_0| \\
 (4.9) \quad &+ ML \int_0^t e^{\omega(t-\tau)} |x(\tau-0) - \hat{x}(\tau-0)| d|\mu|(\tau) \\
 &+ ML \int_0^t e^{\omega(t-\tau)} (1 + |\hat{x}(\tau-0)|) d|\mu - \hat{\mu}|(\tau) \\
 &+ 2ML \int_0^t e^{\omega(t-\tau)} (1 + |x(\tau-0)|) \chi_{\{u \neq \hat{u}\}}(\tau) d|\mu|(\tau).
 \end{aligned}$$

Thus, by Gronwall's inequality, we obtain

$$(4.10) \quad |x(t) - \hat{x}(t)| \leq C[|x_0 - \hat{x}_0| + |\mu - \hat{\mu}|(t) + |\bar{\mu}|(\{u \neq \hat{u}\})], \quad t \in [0, T].$$

Exchanging $(x_0, u(\cdot), \mu(\cdot))$ and $(\hat{x}_0, \hat{u}(\cdot), \hat{\mu}(\cdot))$, we obtain (4.1). \square

The above result gives the continuous dependence of the solution $x(\cdot)$ on the data $(x_0, u(\cdot), \mu(\cdot))$.

Now let us introduce the following:

$$\begin{aligned}
 \tilde{d}((x_0, u(\cdot), \mu(\cdot)), (y_0, v(\cdot), \nu(\cdot))) \\
 (4.11) \quad &= \{|x_0 - y_0|^2 + |\mu - \nu|(T)^2 + |\bar{\mu}|(\{u \neq v\})^2 + |\bar{\nu}|(\{u \neq v\})^2\}^{1/2} \\
 &\quad \forall (x_0, u(\cdot), \mu(\cdot)), (y_0, v(\cdot), \nu(\cdot)) \in X \times \mathcal{U} \times \mathcal{M}.
 \end{aligned}$$

It is clear that \tilde{d} is a metric on the space $X \times \mathcal{U} \times \mathcal{M}$. We have the following lemma.

LEMMA 4.2. *The space $(X \times \mathcal{U} \times \mathcal{M}, \tilde{d})$ is a complete metric space.*

Proof. Let $\{(x_0^n, u^n(\cdot), \mu^n(\cdot))\}$ be a Cauchy sequence in $X \times \mathcal{U} \times \mathcal{M}$ under \tilde{d} . Then we see that

$$(4.12) \quad x_0^n \rightarrow x_0 \quad \text{in } X, \quad (n \rightarrow \infty),$$

for some $x_0 \in X$. Now we show that there exists a $\mu(\cdot) \in \mathcal{M}$ such that

$$(4.13) \quad |\mu^n - \mu|(T) \rightarrow 0, \quad n \rightarrow \infty.$$

In fact, there exists a $\mu(\cdot) \in BV_0([0, T]; K)$ such that (4.13) holds. Thus we must show that $\mu(\cdot) \in \mathcal{M}$. To this end, we let

$$(4.14) \quad \bar{\mu}^n(E) = \int_E \theta_n(\tau) |\bar{\mu}^n|(d\tau) \quad \forall \mathcal{B}([0, T]), \quad n \geq 1$$

and set

$$(4.15) \quad \bar{\nu}^n(E) = \int_E \theta_n(\tau) |\bar{\mu}|(d\tau) \quad \forall \mathcal{B}([0, T]), \quad n \geq 1.$$

Then, by (4.13), we see that

$$(4.16) \quad ||\bar{\mu}^n| - |\bar{\mu}|||(E) \leq |\mu^n - \mu|(T) \rightarrow 0 \quad \forall E \in \mathcal{B}([0, T]), \quad n \rightarrow \infty.$$

Thus, noting $|\theta_n(\tau)| \leq 1$, we obtain

$$(4.17) \quad \begin{aligned} |\bar{\nu}^n - \bar{\mu}|(E) &\leq |\bar{\mu}^n - \bar{\mu}|(E) + \int_E |\theta_n(\tau)| ||\bar{\mu}^n| - |\bar{\mu}|||(d\tau) \\ &\leq 2|\mu^n - \mu|(T) \rightarrow 0 \quad \forall E \in \mathcal{B}([0, T]), \quad n \rightarrow \infty. \end{aligned}$$

Hence we see that

$$(4.18) \quad \int_{[0, T]} |\theta_n(\tau) - \theta_m(\tau)| |\bar{\mu}|(d\tau) = |\bar{\nu}^n - \bar{\nu}^m|([0, T]) \rightarrow 0;$$

i.e., the sequence $\{\theta_n(\cdot)\}$ is Cauchy in $L^1_{|\bar{\mu}|}(0, T)$. Thus there exists a $\theta(\cdot) \in L^1_{|\bar{\mu}|}(0, T)$ such that

$$(4.19) \quad \int_{[0, T]} |\theta_n(\tau) - \theta(\tau)| |\bar{\mu}|(d\tau) \rightarrow 0, \quad n \rightarrow \infty.$$

Then, by (4.15) and (4.17), we obtain

$$(4.20) \quad \bar{\mu}(E) = \int_E \theta(\tau) |\bar{\mu}|(d\tau) \quad \forall E \in \mathcal{B}([0, T]).$$

Then it follows that $\mu(\cdot) \in \mathcal{M}$. Finally, from

$$(4.21) \quad |\bar{\mu}|(\{u^n \neq u^m\}) \leq |\bar{\mu}^n|(\{u^n \neq u^m\}) + |\bar{\mu} - \bar{\mu}^n|([0, T]) \rightarrow 0, \quad m, n \rightarrow \infty,$$

we know that there exists a $u(\cdot)$, such that

$$(4.22) \quad u^n(\cdot) \rightarrow u(\cdot), \quad n \rightarrow \infty, \quad |\bar{\mu}| \text{ a.e.}$$

It is clear that we may assume $u(\cdot)$ to be Borel-measurable. Thus $u(\cdot) \in \mathcal{U}$. Hence the lemma is proved. \square

We should note that in $(X \times \mathcal{U} \times \mathcal{M}, \tilde{d})$,

$$(x_0, u(\cdot), \mu(\cdot)) = (y_0, v(\cdot), \nu(\cdot)),$$

if and only if

$$\begin{aligned} x_0 &= y_0, \\ \mu(\cdot) &= \nu(\cdot), \\ u(t) &= v(t), \quad |\bar{\mu}| \text{ a.e.} \end{aligned}$$

Thus, by changing the values of $u(\cdot)$ on a set of $|\bar{\mu}|$ -measure zero, we have not changed the element $(x_0, u(\cdot), \mu(\cdot))$ in the space $(X \times \mathcal{U} \times \mathcal{M}, \tilde{d})$. On the other hand, from Lemma 4.1, we see that the map from the data $(x_0, u(\cdot), \mu(\cdot))$ to the solution $x(\cdot)$ of (3.3) is well defined from $(X \times \mathcal{U} \times \mathcal{M}, \tilde{d})$ to $D([0, T]; X)$.

LEMMA 4.3. *Let $\lambda(\cdot) \in BV_0([0, T]; \mathbb{R})$ be nondecreasing and $f: \bar{\Delta} \rightarrow X$ be such that, for some nondecreasing function $\omega_f(\cdot): [0, T] \rightarrow \mathbb{R}^+$ with $\omega_f(0) = 0$ and some $|\bar{\lambda}|$ -integrable function $\sigma(\cdot)$, we have that*

$$(4.23) \quad |f(t, \tau) - f(\hat{t}, \tau)| \leq \omega_f(|t - \hat{t}|) \sigma(\tau) \quad \forall (t, \tau), (\hat{t}, \tau) \in \bar{\Delta}.$$

Moreover,

$$(4.24) \quad f(t, \tau) = 0 \quad \text{if } \lambda(\cdot) \text{ jumps at } \tau.$$

Then, for any $\rho \in (0, 1)$, there exists a Borel-measurable set $E_\rho \subset [0, T]$ such that

$$(4.25) \quad \bar{\lambda}(E_\rho) = \rho \bar{\lambda}_c([0, T]) \leq \rho \bar{\lambda}([0, T]),$$

$$(4.26) \quad \rho \int_0^t f(t, \tau) d\lambda(\tau) = \int_0^t f(t, \tau) \chi_{E_\rho}(\tau) d\lambda(\tau) + o(\rho),$$

uniformly in $t \in [0, T]$.

Proof. By (4.24), we see that

$$(4.27) \quad \int_0^t f(t, \tau) d\lambda(\tau) = \int_{[0, T]} f(t, \tau) \bar{\lambda}_c(d\tau).$$

Then, by the arguments used to prove the similar result for $\bar{\lambda}_c(\cdot)$ being Lebesgue-measurable (see [22], [23]), we can obtain a Borel-measurable set $E_\rho \subset [0, T]$ such that

$$(4.28) \quad \bar{\lambda}_c(E_\rho) = \rho \bar{\lambda}_c([0, T]),$$

and (4.26) holds. Since the jump points of $\lambda(\cdot)$ is at most a countable set, we may change E_ρ somewhat so that

$$(4.29) \quad \bar{\lambda}_c(E_\rho) = \bar{\lambda}(E_\rho).$$

Hence (4.25) follows. \square

COROLLARY 4.4. *Let $\mu(\cdot) \in \mathcal{M}$ and $G(t, \tau): \bar{\Delta} \rightarrow \mathcal{L}(Z, X)$ be uniformly Borel-measurable satisfying the following:*

$$(4.30) \quad \|G(t, \tau)\|_{\mathcal{L}(Z, X)} \leq a(\tau) \quad \forall (t, \tau) \in \bar{\Delta},$$

$$(4.31) \quad \|G(t, \tau) - G(s, \tau)\|_{\mathcal{L}(Z, X)} \leq \omega_G(|t - s|) b(\tau) \quad \forall (t, \tau), (s, \tau) \in \bar{\Delta},$$

where $a(\cdot), b(\cdot) \in B_{|\bar{\mu}|}([0, T]; \mathbb{R})$, and $\omega_G(\cdot): [0, T] \rightarrow \mathbb{R}^+$ is a nondecreasing function with $\omega_G(0) = 0$. Moreover, we let

$$(4.32) \quad G(t, \tau) = 0 \quad \text{if } \mu(\cdot) \text{ jumps at } \tau.$$

Then, for any $\rho \in (0, 1)$, there exists a Borel-measurable set $E_\rho \subset [0, T]$ such that

$$(4.33) \quad |\bar{\mu}|(E_\rho) = \rho |\bar{\mu}_c|([0, T]) \leq \rho |\mu|(T),$$

$$(4.34) \quad \rho \int_0^t G(t, \tau) d\mu(\tau) = \int_0^t G(t, \tau) \chi_{E_\rho}(\tau) d\mu(\tau) + o(\rho),$$

uniformly in $t \in [0, T]$.

LEMMA 4.5. Let $(u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)), (u^\#(\cdot), \mu^\#(\cdot)) \in \mathcal{U} \times \mathcal{M}$ be such that

$$(4.35) \quad \lim_{\varepsilon \rightarrow 0} |\mu^\varepsilon - \mu^\#|(T) = 0,$$

$$(4.36) \quad \lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\#|(\{u^\varepsilon \neq u^\#\}) = 0, \quad \lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\varepsilon|(\{u^\varepsilon \neq u^\#\}) = 0.$$

Let

$$D_\varepsilon = \{\tau \in [0, T] \mid \mu^\varepsilon(\cdot) \text{ jumps at } \tau\},$$

$$D_\# = \{\tau \in [0, T] \mid \mu^\#(\cdot) \text{ jumps at } \tau\}.$$

For $u(\cdot) \in \mathcal{U}$, with

$$(4.37) \quad u(\tau) = u^\#(\tau), \quad \tau \in D_\#,$$

set

$$(4.38) \quad u_\varepsilon(\cdot) = u(\cdot)\chi_{[0, T] \setminus D_\varepsilon}(\cdot) + u^\varepsilon(\cdot)\chi_{D_\varepsilon}(\cdot).$$

Then

$$(4.39) \quad \lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\#|(\{u_\varepsilon \neq u\}) = 0, \quad \lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\varepsilon|(\{u^\varepsilon \neq u\}) = 0.$$

Proof. For $n \geq 1$, let

$$D_\varepsilon^n = \{\tau \in [0, T] \mid |\mu^\varepsilon(\tau+0) - \mu^\varepsilon(\tau)| \text{ or } |\mu^\varepsilon(\tau) - \mu^\varepsilon(\tau-0)| \geq 1/n\},$$

$$D_\#^n = \{\tau \in [0, T] \mid |\mu^\#(\tau+0) - \mu^\#(\tau)| \text{ or } |\mu^\#(\tau) - \mu^\#(\tau-0)| \geq 1/n\}.$$

Then, by (4.35), we see that, for any $n \geq 1$, there exists an $\varepsilon_n > 0$, such that, for all $\varepsilon \in (0, \varepsilon_n)$,

$$(4.40) \quad D_\varepsilon^n = D_\#^n.$$

Thus, by noting the fact that

$$D_\varepsilon = \bigcup_{n \geq 1} D_\varepsilon^n, \quad D_\# = \bigcup_{n \geq 1} D_\#^n,$$

we have that

$$(4.41) \quad \lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\varepsilon|(D_\varepsilon \setminus D_\#) = 0, \quad \lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\varepsilon|(D_\# \setminus D_\varepsilon) = 0,$$

$$\lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\#|(D_\varepsilon \setminus D_\#) = 0, \quad \lim_{\varepsilon \rightarrow 0} |\bar{\mu}^\#|(D_\# \setminus D_\varepsilon) = 0.$$

Next, by (4.37), we have that

$$(4.42) \quad u(\cdot) = u(\cdot)\chi_{[0, T] \setminus D_\#}(\cdot) + u^\#(\cdot)\chi_{D_\#}(\cdot).$$

Thus we obtain

$$(4.43) \quad \{u_\varepsilon \neq u\} \subseteq D_\varepsilon \cup D_\#,$$

while

$$D_\varepsilon \cup D_\# = (D_\varepsilon \setminus D_\#) \cup (D_\# \setminus D_\varepsilon) \cup (D_\varepsilon \cap D_\#)$$

and

$$(D_\varepsilon \cap D_\#) \cap \{u_\varepsilon \neq u\} \subseteq \{u^\varepsilon \neq u^\#\}.$$

Hence (4.39) follows from (4.36) and (4.41). \square

5. Proof of maximum principle. In this section, we give a proof of the maximum principle stated in § 3. As assumed in § 3, we let $(x^\#(\cdot), u^\#(\cdot), \mu^\#(\cdot))$ be an optimal solution of Problem C such that the condition of Theorem 3.2 concerning $\mathcal{R} - \mathcal{Q}$ holds. For any $(x_0, u(\cdot), \mu(\cdot)) \in X \times \mathcal{U} \times \mathcal{M}$, we let $x(\cdot; x_0, u(\cdot), \mu(\cdot))$ be the corresponding solution of (3.3), and we define

$$(5.1) \quad x^0(t) = \int_0^t \langle f(\tau, x(\tau; x_0, u(\cdot), \mu(\cdot)), u(\tau)), d\mu(\tau) \rangle, \quad t \in [0, T].$$

Then we see that $x^0(T) = J(x(\cdot), u(\cdot), \mu(\cdot))$, the cost functional of Problem C. Now, for any $\varepsilon > 0$, we define

$$(5.2) \quad \begin{aligned} J_\varepsilon(x_0, u(\cdot), \mu(\cdot)) &= \{d_\Omega(x_0, x(T; x_0, u(\cdot), \mu(\cdot)))^2 \\ &\quad + d_{\Omega^0(\varepsilon)}(x^0(T; x_0, u(\cdot), \mu(\cdot)))^2\}^{1/2} \\ \forall(x_0, u(\cdot), \mu(\cdot)) &\in X \times \mathcal{U} \times \mathcal{M}, \end{aligned}$$

where

$$\begin{aligned} \Omega^0(\varepsilon) &= (-\infty, -\varepsilon + x^{\#0}(T)], \\ d_\Omega(x_0, x_1) &= \inf \{(|x_0 - y_0|^2 + |x_1 - y_1|^2)^{1/2} \mid (y_0, y_1) \in \Omega\}, \\ d_{\Omega^0(\varepsilon)}(x^0) &= \inf \{|x^0 - y^0| \mid y^0 \in \Omega^0(\varepsilon)\}. \end{aligned}$$

Then it is clear that

$$(5.3) \quad J_\varepsilon(x_0, u(\cdot), \mu(\cdot)) > 0 \quad \forall(x_0, u(\cdot), \mu(\cdot)) \in X \times \mathcal{U} \times \mathcal{M},$$

$$(5.4) \quad J_\varepsilon(x^\#(0), u^\#(\cdot), \mu^\#(\cdot)) = \varepsilon \leq \inf_{X \times \mathcal{U} \times \mathcal{M}} J_\varepsilon(x_0, u(\cdot), \mu(\cdot)) + \varepsilon.$$

Thus by Ekeland's variational principle [15], [16], we can find an $(x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)) \in X \times \mathcal{U} \times \mathcal{M}$, such that

$$(5.5) \quad \tilde{d}((x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)), (x^\#(0), u^\#(\cdot), \mu^\#(\cdot))) \leq \sqrt{\varepsilon},$$

$$(5.6) \quad J_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)) \leq J_\varepsilon(x^\#(0), u^\#(\cdot), \mu^\#(\cdot)),$$

$$(5.7) \quad \begin{aligned} J_\varepsilon(x_0, u(\cdot), \mu(\cdot)) &\geq J_\varepsilon(x^\varepsilon(0), u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)) \\ &\quad - \sqrt{\varepsilon} \tilde{d}((x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)), (x_0, u(\cdot), \mu(\cdot))) \\ &\quad \forall(x_0, u(\cdot), \mu(\cdot)) \in X \times \mathcal{U} \times \mathcal{M}. \end{aligned}$$

We let $x^\varepsilon(\cdot)$ be the solution of (3.3) corresponding to $(x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot))$ and $x^{0,\varepsilon}(\cdot)$ be the corresponding function defined by (5.1). Next, we let $D_\#$ and D_ε be the same as those defined in § 4 and fix any $\rho \in (0, 1)$ and $(x_0, u(\cdot), \mu(\cdot)) \in X \times \mathcal{U} \times \mathcal{M}$ with

$$(5.8) \quad u(\cdot) \in \mathcal{U}^\# \equiv \{u(\cdot) \in \mathcal{U} \mid u(\tau) = u^\#(\tau), \forall \tau \in D_\#\}.$$

Then let

$$(5.9) \quad u_\varepsilon(\tau) = u(\tau)\chi_{[0, T] \setminus D_\varepsilon}(\tau) + u^\varepsilon(\tau)\chi_{D_\varepsilon}(\tau), \quad \tau \in [0, T].$$

By Lemma 4.5, we have that

$$(5.10) \quad |\bar{\mu}^\#|(\{u_\varepsilon \neq u\}), \quad |\bar{\mu}^\varepsilon|(\{u_\varepsilon \neq u\}) \rightarrow 0, \quad \varepsilon \rightarrow 0.$$

Next, by Corollary 4.4, we can find a Borel-measurable set $E_\rho \subset [0, T]$ such that

$$(5.11) \quad |\bar{\mu}^\varepsilon|(E_\rho) = \rho |\bar{\mu}_0^\varepsilon|([0, T]) \leq \rho |\bar{\mu}^\varepsilon|([0, T])$$

and

$$(5.12) \quad \rho \int_0^t e^{A(t-\tau)} \Delta F^\varepsilon(\tau) d\mu^\varepsilon(\tau) = \int_0^t e^{A(t-\tau)} \Delta F^\varepsilon(\tau) \chi_{E_\rho}(\tau) d\mu^\varepsilon(\tau) + o(\rho),$$

uniformly in $t \in [0, T]$, where

$$\Delta F^\varepsilon(\tau) = F(\tau, x^\varepsilon(\tau-0), u_\varepsilon(\tau)) - F(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau)), \quad \tau \in [0, T].$$

Then we define

$$(5.13) \quad \begin{aligned} x_{0,\rho}^\varepsilon &= x_0^\varepsilon + \rho x_0, \\ u_\rho^\varepsilon(\cdot) &= u^\varepsilon(\cdot) \chi_{[0,T] \setminus E_\rho}(\cdot) + u_\varepsilon(\cdot) \chi_{E_\rho}(\cdot), \\ \mu_\rho^\varepsilon(\cdot) &= \mu^\varepsilon(\cdot) + \rho(\mu(\cdot) - \mu^\varepsilon(\cdot)). \end{aligned}$$

It is clear that $(x_{0,\rho}^\varepsilon, u_\rho^\varepsilon(\cdot), \mu_\rho^\varepsilon(\cdot)) \in X \times \mathcal{U} \times \mathcal{M}$. Now let $x_\rho^\varepsilon(\cdot)$ be the unique solution of (3.3) corresponding to $(x_{0,\rho}^\varepsilon, u_\rho^\varepsilon(\cdot), \mu_\rho^\varepsilon(\cdot))$ and $x_\rho^{\varepsilon,0}(\cdot)$ be defined as in (5.1) correspondingly. We now would like to derive the variation equation associated with (3.3). We observe the following:

$$(5.14) \quad \begin{aligned} & \frac{1}{\rho} [x_\rho^\varepsilon(t) - x^\varepsilon(t)] \\ &= e^{At} x_0 + \int_0^t e^{A(t-\tau)} F(\tau, x_\rho^\varepsilon(\tau-0), u_\rho^\varepsilon(\tau)) d(\mu(\tau) - \mu^\varepsilon(\tau)) \\ & \quad + \frac{1}{\rho} \int_0^t e^{A(t-\tau)} [F(\tau, x_\rho^\varepsilon(\tau-0), u_\rho^\varepsilon(\tau)) - F(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau))] d\mu^\varepsilon(\tau) \\ & \quad + \frac{1}{\rho} \int_0^t e^{A(t-\tau)} [F(\tau, x^\varepsilon(\tau-0), u(\tau)) - F(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau))] \chi_{E_\rho}(\tau) d\mu^\varepsilon(\tau). \end{aligned}$$

By (5.13) and Lemma 4.1, we see that

$$(5.15) \quad \lim_{\rho \rightarrow 0} \sup_{0 \leq t \leq T} |x_\rho^\varepsilon(t) - x^\varepsilon(t)| = 0.$$

Hence a careful calculation shows that (see (5.12))

$$(5.16) \quad \begin{aligned} & \frac{1}{\rho} [x_\rho^\varepsilon(t) - x^\varepsilon(t)] \\ &= e^{At} x_0 + \int_0^t e^{A(t-\tau)} F(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau)) (d\mu(\tau) - \mu^\varepsilon(\tau)) \\ & \quad + \int_0^t e^{A(t-\tau)} F_x \left(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau); \frac{x_\rho^\varepsilon(\tau-0) - x^\varepsilon(\tau-0)}{\rho} \right) d\mu^\varepsilon(\tau) \\ & \quad + \int_0^t e^{A(t-\tau)} \Delta F^\varepsilon(\tau) d\mu^\varepsilon(\tau) + o(1), \end{aligned}$$

uniformly in $t \in [0, T]$, where

$$(5.17) \quad F_x(\tau, x, u; \xi) = \lim_{h \rightarrow 0} \frac{F(\tau, x + h\xi, u) - F(\tau, x, u)}{h},$$

the Fréchet derivative of $F(\tau, x, u)$ in x in the direction ξ . Hence the limit

$$(5.18) \quad \lim_{\rho \rightarrow 0} \frac{x_\rho^\varepsilon(\tau) - x^\varepsilon(t)}{\rho} = \xi_\varepsilon(t), \quad t \in [0, T],$$

exists, and $\xi_\varepsilon(\cdot)$ satisfies

$$(5.19) \quad \begin{aligned} \xi_\varepsilon(t) = & e^{At}x_0 + \int_0^t e^{A(t-\tau)} F_x(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau); \xi_\varepsilon(\tau-0)) d\mu^\varepsilon(\tau) \\ & + \int_0^t e^{A(t-\tau)} F(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau)) d(\mu(\tau) - \mu^\varepsilon(\tau)) \\ & + \int_0^t e^{A(t-\tau)} \Delta F^\varepsilon(\tau) d\mu^\varepsilon(\tau), \quad t \in [0, T]. \end{aligned}$$

Similarly, we have the existence of the limit

$$(5.20) \quad \lim_{\rho \rightarrow 0} \frac{x_\rho^{0,\varepsilon}(\tau) - x^{0,\varepsilon}(t)}{\rho} = \xi_\varepsilon^0(t), \quad t \in [0, T],$$

and $\xi_\varepsilon^0(\cdot)$ satisfies

$$(5.21) \quad \begin{aligned} \xi_\varepsilon^0(t) = & \int_0^t \langle f_x(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau)) \xi_\varepsilon(\tau-0), d\mu^\varepsilon(\tau) \rangle \\ & + \int_0^t \langle f(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau)), d(\mu(\tau) - \mu^\varepsilon(\tau)) \rangle \\ & + \int_0^t \langle \Delta f^\varepsilon(\tau), d\mu^\varepsilon(\tau) \rangle, \quad t \in [0, T], \end{aligned}$$

where

$$\Delta f^\varepsilon(\tau) = f(\tau, x^\varepsilon(\tau-0), u_\varepsilon(\tau)) - f(\tau, x^\varepsilon(\tau-0), u^\varepsilon(\tau)), \quad \tau \in [0, T].$$

Then, from (5.5), we know that

$$(5.22) \quad \begin{aligned} \lim_{\varepsilon \rightarrow 0} x^\varepsilon(t) &= x^\#(t), & \lim_{\varepsilon \rightarrow 0} x_\varepsilon^0(t) &= x^{\#,0}(t), & t \in [0, T], \\ \lim_{\varepsilon \rightarrow 0} \xi_\varepsilon(t) &= \xi(t), & \lim_{\varepsilon \rightarrow 0} \xi_\varepsilon^0(t) &= \xi^0(t), & t \in [0, T], \end{aligned}$$

with

$$(5.23) \quad \begin{aligned} \xi(t) = & e^{At}x_0 + \int_0^t e^{A(t-\tau)} F_x(\tau, x^\#(\tau-0), u^\#(\tau); \xi(\tau-0)) d\mu^\#(\tau) \\ & + \int_0^t e^{A(t-\tau)} F(\tau, x^\#(\tau-0), u^\#(\tau)) d(\mu(\tau) - \mu^\#(\tau)) \\ & + \int_0^t e^{A(t-\tau)} \Delta F^\#(\tau) d\mu^\#(\tau), \quad t \in [0, T]; \end{aligned}$$

$$(5.24) \quad \begin{aligned} \xi^0(t) = & \int_0^t \langle f_x(\tau, x^\#(\tau-0), u^\#(\tau)) \xi(\tau-0), d\mu^\#(\tau) \rangle \\ & + \int_0^t \langle f(\tau, x^\#(\tau-0), u^\#(\tau)), d(\mu(\tau) - \mu^\#(\tau)) \rangle \\ & + \int_0^t \langle \Delta f^\#(\tau), d\mu^\#(\tau) \rangle, \quad t \in [0, T], \end{aligned}$$

where

$$\begin{aligned}\Delta F^\#(\tau) &= F(\tau, x^\#(\tau-0), u(\tau)) - F(\tau, x^\#(\tau-0), u^\#(\tau)), & \tau \in [0, T], \\ \Delta f^\#(\tau) &= f(\tau, x^\#(\tau-0), u(\tau)) - f(\tau, x^\#(\tau-0), u^\#(\tau)), & \tau \in [0, T].\end{aligned}$$

Next, by (5.7), we have that

$$\begin{aligned}(5.25) \quad & -\sqrt{\varepsilon} \tilde{d}((x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)), (x_{0,\rho}^\varepsilon, u_\rho^\varepsilon(\cdot), \mu_\rho^\varepsilon(\cdot))) \\ & \leq J_\varepsilon(x_{0,\rho}^\varepsilon, u_\rho^\varepsilon(\cdot), \mu_\rho^\varepsilon(\cdot)) - J_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot)).\end{aligned}$$

Then, by (5.11), (5.13), and (5.15)–(5.21), after dividing ρ and sending $\rho \rightarrow 0$ (similar to [23]), we have that

$$\begin{aligned}(5.26) \quad & -\sqrt{\varepsilon} \{ |x_0|^2 + |\mu^\varepsilon|(T)^2 + [|\mu^\varepsilon|(T) + |\mu|(T)]^2 \}^{1/2} \\ & \leq \frac{1}{J_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot))} \{ d_\Omega^0((x_0^\varepsilon, x^\varepsilon(T)); (x_0, \xi_\varepsilon(T))) d_\Omega(x_0^\varepsilon, x^\varepsilon(T)) \\ & \quad + d_{\Omega^0(\varepsilon)}^0(x^{0,\varepsilon}(T); \xi_\varepsilon^0(T)) d_{\Omega^0(\varepsilon)}(x^{0,\varepsilon}(T)) \} \\ & = \langle \bar{\varphi}_\varepsilon, x_0 \rangle + \langle \bar{\psi}_\varepsilon, \xi_\varepsilon(T) \rangle + \bar{\psi}_\varepsilon^0 \xi_\varepsilon^0(T),\end{aligned}$$

where

$$(5.27) \quad \bar{\varphi}_\varepsilon = \frac{d_\Omega(x_0^\varepsilon, x^\varepsilon(T)) a_\varepsilon}{J_\varepsilon(x_0, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot))},$$

$$(5.28) \quad \bar{\psi}_\varepsilon = \frac{d_\Omega(x_0^\varepsilon, x^\varepsilon(T)) b_\varepsilon}{J_\varepsilon(x_0, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot))},$$

$$(5.29) \quad \bar{\psi}_\varepsilon^0 = \frac{d_{\Omega^0(\varepsilon)}(x^{0,\varepsilon}(T))}{J_\varepsilon(x_0, u^\varepsilon(\cdot), \mu^\varepsilon(\cdot))},$$

with

$$(5.30) \quad \{(a_\varepsilon, b_\varepsilon)\} = \partial d_\Omega(x_0^\varepsilon, x^\varepsilon(T)).$$

Here the notations of Clarke's generalized directional derivative and gradient of functions d_Ω and $d_{\Omega^0(\varepsilon)}$ are adopted (see [10] for details). By Hypothesis 1, as in [23], we know that $\partial d_\Omega(x_0^\varepsilon, x^\varepsilon(T))$ is a singleton. Also, we have that (see [23])

$$(5.31) \quad |\bar{\varphi}_\varepsilon|_{X^*}^2 + |\bar{\psi}_\varepsilon|_{X^*}^2 + (\bar{\psi}_\varepsilon^0)^2 = 1 \quad \forall \varepsilon > 0.$$

Since Ω is convex, by (5.30), we have that

$$(5.32) \quad \langle a_\varepsilon, y_0 - x_0^\varepsilon \rangle + \langle b_\varepsilon, y_1 - x^\varepsilon(T) \rangle \leq 0 \quad \forall (y_0, y_1) \in \Omega.$$

Thus

$$\begin{aligned}(5.33) \quad & \langle \bar{\varphi}_\varepsilon, y_0 - x^\#(0) \rangle + \langle \bar{\psi}_\varepsilon, y_1 - x^\#(T) \rangle \\ & \leq \{ |x_0^\varepsilon - x_0^\#|^2 + |x^\varepsilon(T) - x^\#(T)|^2 \}^{1/2} \equiv \delta_\varepsilon \rightarrow 0, \quad (\varepsilon \rightarrow 0).\end{aligned}$$

Then, by (5.22) and (5.26), we obtain

$$\begin{aligned}(5.34) \quad & \langle \bar{\varphi}_\varepsilon, x_0 - (y_0 - x^\#(0)) \rangle + \langle \bar{\psi}_\varepsilon, \xi(T) - (y_1 - x^\#(T)) \rangle + \bar{\psi}_\varepsilon^0 \xi_\varepsilon^0(T) \geq -\hat{\delta}_\varepsilon \\ & \quad \forall (y_0, y_1) \in \Omega,\end{aligned}$$

with

$$(5.35) \quad \lim_{\varepsilon \rightarrow 0} \hat{\delta}_\varepsilon = 0,$$

uniformly in $(x_0, u(\cdot), \mu(\cdot))$ in bounded sets of $X \times \mathcal{U} \times \mathcal{M}$. Next, let (see (3.7))

$$(5.36) \quad B(\tau, \xi, z) = F_x(\tau, x^\#(\tau-0), u^\#(\tau); \xi)z \quad \forall (\tau, \xi, z) \in [0, T] \times X \times Z,$$

$$(5.37) \quad \nu(\tau) = \begin{pmatrix} \mu(\tau) \\ \mu^\#(\tau) \end{pmatrix} \quad \forall \tau \in [0, T],$$

and

$$(5.38) \quad \bar{G}(\tau) = (F(\tau, x^\#(\tau-0), u^\#(\tau)), \Delta F^\#(\tau)) \quad \forall \tau \in [0, T].$$

Then (5.23) reads

$$(5.39) \quad \begin{aligned} \xi(t) = & e^{A't}x_0 + \int_0^t e^{A(t-\tau)} B(\tau, \xi(\tau-0), d\mu^\#(\tau)) \\ & + \int_0^t e^{A(t-\tau)} \bar{G}(\tau) d\nu(\tau) \quad \forall t \in [0, T]. \end{aligned}$$

Similarly, (5.24) can be written as

$$(5.40) \quad \xi^0(t) = \int_0^t \langle f_x(\tau) \xi(\tau-0), d\mu^\#(\tau) \rangle + \int_0^t \langle \bar{f}(\tau), d\nu(\tau) \rangle \quad \forall t \in [0, T],$$

with

$$\begin{aligned} f_x(\tau) &= f_x(\tau, x^\#(\tau-0), u^\#(\tau)), \quad \tau \in [0, T], \\ \bar{f}(\tau) &= (f(\tau, x^\#(\tau-0), u^\#(\tau)), \Delta f^\#(\tau)), \quad \tau \in [0, T]. \end{aligned}$$

By Theorem 2.6, we obtain

$$(5.41) \quad \xi(t) = \Phi(t, 0)x_0 + \int_0^t \Psi(t, \tau) \bar{G}(\tau) d\nu(\tau), \quad t \in [0, T].$$

Then we see that

$$(5.42) \quad \mathcal{R} = \left\{ \int_0^t \Psi(t, \tau) \bar{G}(\tau) d\nu(\tau) \mid (u(\cdot), \mu(\cdot)) \in \mathcal{U}^\# \times \mathcal{M} \right\}.$$

Since we assume that $\mathcal{R} - \mathcal{Q}$ is finite-codimensional in X , it is easy to see that (see [23]) the set

$$\left\{ \left(\Phi(T, 0)x_0 + \xi \right) \mid \xi \in \mathcal{R}, \eta \in X \right\} - \Omega$$

is finite-codimensional in $X \times X$. Thus, as [17], [23], from (5.31) and (5.34), we can find a subsequence $\varepsilon \downarrow 0$, such that

$$(5.43) \quad (\bar{\varphi}_\varepsilon, \bar{\psi}_\varepsilon, \bar{\psi}_\varepsilon^0) \xrightarrow{*} (\bar{\varphi}, \bar{\psi}, \bar{\psi}^0) \neq 0.$$

Then, from (5.22), (5.26), and (5.29), we have that

$$(5.44) \quad \langle \bar{\varphi}, x_0 \rangle + \langle \bar{\psi}, \xi(T) \rangle + \bar{\psi}^0 \xi^0(T) \geq 0,$$

$$(5.45) \quad \bar{\psi}^0 \geq 0.$$

From (5.32), we have that

$$(5.46) \quad \langle \bar{\varphi}, y_0 - x_0^\# \rangle + \langle \bar{\psi}, y_1 - x^\#(T) \rangle \leq 0 \quad \forall (y_0, y_1) \in \Omega.$$

Next, we define

$$(5.47) \quad \psi^0 = -\bar{\psi}^0 \leq 0$$

and

$$(5.48) \quad \begin{aligned} \psi(t) &= \Psi(T, t)^* \psi(T) + \psi^0 \int_t^T \Psi(\tau - 0, t)^* f_x(\tau)^* d\mu^\#(\tau), \quad t \in [0, T], \\ \psi(T) &= -\bar{\psi}. \end{aligned}$$

Then we have that

$$(5.49) \quad \begin{aligned} 0 &\geq \langle \psi(T), \xi(T) \rangle - \langle \psi(0), \xi(0) \rangle + \psi^0 \xi^0(T) + \langle \psi(0) - \bar{\varphi}, x_0 \rangle \\ &= \left\langle \psi(T), \Phi(T, 0)x_0 + \int_0^T \Psi(T, \tau) \bar{G}(\tau) d\nu(\tau) \right\rangle \\ &\quad - \left\langle \Psi(T, 0)^* \psi(T) + \psi^0 \int_0^T \Psi(\tau - 0, 0)^* f_x(\tau)^* d\mu^\#(\tau), x_0 \right\rangle \\ &\quad + \psi^0 \int_0^T \left\langle f_x(\tau) \left[\Phi(\tau - 0, 0)x_0 + \int_{[0, \tau)} \Psi(\tau - 0, s) \bar{G}(s) \bar{\nu}(ds) \right], d\mu^\#(\tau) \right\rangle \\ &\quad + \psi^0 \int_0^T \langle \bar{f}(\tau), d\nu(\tau) \rangle + \langle \psi(0) - \bar{\varphi}, x_0 \rangle \\ &= \int_0^T \langle \Psi(T, \tau)^* \psi(T), \bar{G}(\tau) d\nu(\tau) \rangle - \langle \Theta(T, 0)^* \psi(T), x_0 \rangle \\ &\quad + \psi^0 \int_0^T \int_0^\tau \langle \Psi(\tau - 0, s)^* f_x(\tau)^* d\mu^\#(\tau), \bar{G}(s) d\nu(s) \rangle \\ &\quad + \psi^0 \int_0^T \langle \bar{f}(\tau), d\nu(\tau) \rangle + \langle \psi(0) - \bar{\varphi}, x_0 \rangle \\ &= \int_0^T \langle \psi(\tau), \bar{G}(\tau) d\nu(\tau) \rangle + \int_0^T \langle \psi^0 \bar{f}(\tau), d\nu(\tau) \rangle \\ &\quad + \langle \psi(0) - \bar{\varphi} - \Theta(T, 0)^* \psi(T), x_0 \rangle. \end{aligned}$$

The above holds for all $(x_0, u(\cdot), \mu(\cdot)) \in X \times \mathcal{U}^\# \times \mathcal{M}$. Thus, by taking $x_0 = 0$, we obtain

$$(5.50) \quad \int_0^T \langle \psi(\tau), \bar{G}(\tau) d\nu(\tau) \rangle + \int_0^T \langle \psi^0 \bar{f}(\tau), d\nu(\tau) \rangle \leq 0,$$

i.e.,

$$(5.51) \quad \begin{aligned} 0 &\geq \int_0^T \langle H(\tau, x^\#(\tau - 0), u^\#(\tau), \psi^0, \psi(\tau)), d(\mu(\tau) - \mu^\#(\tau)) \rangle \\ &\quad + \int_0^T \langle H(\tau, x^\#(\tau - 0), u(\tau), \psi^0, \psi(\tau)) \\ &\quad \quad - H(\tau, x^\#(\tau - 0), u^\#(\tau), \psi^0, \psi(\tau)), d\mu^\#(\tau) \rangle \end{aligned}$$

$$\forall (u(\cdot), \mu(\cdot)) \in \mathcal{U}^\# \times \mathcal{M}.$$

Hence (3.15) and (3.16) follow. Now, by taking $(u(\cdot), \mu(\cdot)) = (u^\#(\cdot), 0)$ in (5.49), we obtain

$$(5.52) \quad \langle \psi(0) - \bar{\varphi} - \Theta(T, 0)^* \psi(T), x_0 \rangle \leq 0 \quad \forall x_0 \in X.$$

Thus

$$(5.53) \quad \bar{\varphi} = \psi(0) - \Theta(T, 0)^* \psi(T).$$

Then (5.46) reads

$$(5.54) \quad \langle \psi(0) - \Theta(T, 0)^* \psi(T), y_0 - x_0^\# \rangle - \langle \psi(T), y_1 - x^\#(T) \rangle \leq 0 \quad \forall (y_0, y_1) \in \Omega.$$

This gives the transversality condition (3.17). By (5.43), (5.47), (5.48), and (5.53), we see that

$$(5.55) \quad (\psi(\cdot), \psi^0) \neq 0.$$

Finally, by (5.48) and Lemma 2.7, we have that

$$(5.56) \quad \begin{aligned} \psi(t) &= e^{A^*(T-t)} \psi(T) + \int_t^T e^{A^*(\tau-t)} B(\tau, d\mu^\#(\tau))^* \Psi(T, \tau)^* \psi(T) \\ &\quad + \int_t^T \psi^0 e^{A^*(\tau-t)} f_x(\tau)^* d\mu^\#(\tau) \\ &\quad + \int_t^T \psi^0 \int_t^\tau e^{A^*(s-t)} B(s, d\mu^\#(s))^* \Psi(\tau - 0, s)^* f_x(\tau)^* d\mu^\#(\tau) \\ &= e^{A^*(T-t)} \psi(T) + \int_t^T e^{A^*(\tau-t)} B(\tau, d\mu^\#(\tau))^* \psi(T) \\ &\quad + \int_t^T e^{A^*(\tau-t)} \psi^0 f_x(\tau)^* d\mu^\#(\tau). \end{aligned}$$

Thus, the function $\psi(\cdot)$ defined by (5.48) is the solution of (3.12). Hence we have completed the proof of Theorem 3.2. \square

Appendix. In this appendix, we present some relevant results concerning the Young integral and the vector measures, among others.

First, we let $T > 0$ be a constant, Z be a Banach space, and K be a subset of Z . We define

$$BV([0, T]; K) = \{\mu(\cdot) : [0, T] \rightarrow K \mid \mu(\cdot) \text{ is of bounded variation}\},$$

$$BV_0([0, T]; K) = \{\mu(\cdot) \in BV([0, T]; K) \mid \mu(0) = 0\}.$$

For convenience, we take the convention that, for any $\mu(\cdot) \in BV([0, T]; K)$,

$$(A.1) \quad \mu(0 - 0) = \mu(0), \quad \mu(T + 0) = \mu(T).$$

Next, for any $\mu(\cdot) \in BV([0, T]; K)$, we define

$$(A.2) \quad \begin{aligned} \bar{\mu}([0, t]) &= \mu(t + 0) - \mu(0), & 0 \leq t \leq T, \\ \bar{\mu}((s, t]) &= \mu(t + 0) - \mu(s + 0), & 0 \leq s \leq t \leq T. \end{aligned}$$

Extending this vector-valued set function to $\mathcal{B}([0, T])$, the Borel σ -field of $[0, T]$, we obtain a vector measure $\bar{\mu}(\cdot)$, which is of bounded variation. We refer to $\bar{\mu}(\cdot)$ as the

vector measure induced by $\mu(\cdot)$. For $\mu(\cdot) \in BV([0, T]; K)$, we let $|\mu|(\cdot) \in BV([0, T]; \mathbb{R}^+)$ be the variation of $\mu(\cdot)$; i.e., $|\mu|(t)$ is the total variation of $\mu(\cdot)$ on the interval $[0, t]$. It is clear that

$$(A.3) \quad \begin{cases} |\bar{\mu}|([0, t]) = |\mu|(t+0), \\ |\bar{\mu}|([0, t)) = |\mu|(t-0), \end{cases} \quad \forall 0 \leq t \leq T,$$

where $|\bar{\mu}|(\cdot)$ is the variation of the vector measure $\bar{\mu}(\cdot)$ (see [13]). Next, by noting that any $\mu(\cdot) \in BV([0, T]; K)$ has at most countably many discontinuity points, we may define

$$(A.4) \quad \begin{aligned} \mu_b(t) &= \sum_{0 \leq \tau < t} [\mu(\tau+0) - \mu(\tau-0)] + \mu(t) - \mu(t-0) \quad \forall t \in [0, T], \\ \mu_c(t) &= \mu(t) - \mu_b(t) \quad \forall t \in [0, T]. \end{aligned}$$

Then, it is easy to show (mimic the proof for the scalar-valued case; see [27]) that $\mu_c(\cdot) \in BV([0, T]; Z) \cap C([0, T]; Z)$ and

$$(A.5) \quad |\mu|(t) = |\mu_b|(t) + |\mu_c|(t) \quad \forall t \in [0, T].$$

Now we let $\bar{\mu}_b(\cdot)$ and $\bar{\mu}_c(\cdot)$ be the vector measures induced by $\mu_b(\cdot)$ and $\mu_c(\cdot)$, respectively. Then we have that

$$(A.6) \quad |\bar{\mu}|(E) = |\bar{\mu}_b|(E) + |\bar{\mu}_c|(E) \quad \forall E \in \mathcal{B}([0, T]).$$

Next, if Z has the RNP (Radon-Nikodym property [13]) with respect to $|\bar{\mu}|$, then there exists a $|\bar{\mu}|$ -integrable function $\theta(\cdot): [0, T] \rightarrow Z$ such that

$$(A.7) \quad \bar{\mu}(E) = \int_E \theta(\tau) |\bar{\mu}|(d\tau) \quad \forall E \in \mathcal{B}([0, T]),$$

$$(A.8) \quad |\bar{\mu}|(E) = \int_E |\theta(\tau)|_Z |\bar{\mu}|(d\tau) \quad \forall E \in \mathcal{B}([0, T])$$

(see [13] for a proof). Thus

$$(A.9) \quad |\theta(\tau)|_Z = 1, \quad |\bar{\mu}| \text{ a.e. } \tau \in [0, T].$$

Let us set

$$(A.10) \quad \begin{aligned} \mathcal{M}_0([0, T]; K) &= \left\{ \mu(\cdot) \in BV_0([0, T]; K) \mid \bar{\mu}(E) = \int_E \theta(\tau) |\bar{\mu}|(d\tau), \right. \\ &\quad \left. \forall E \in \mathcal{B}([0, T]), \text{ for some } \theta(\cdot), \text{ with } |\theta(\tau)|_Z = 1, |\bar{\mu}| \text{ a.e.} \right\}. \end{aligned}$$

PROPOSITION A.1. *Let K be a convex cone in Z . Then $\mathcal{M}_0 \equiv \mathcal{M}_0([0, T]; K)$ is a convex cone.*

The proof is straightforward.

Now let X be another Banach space. An operator-valued function $F(\cdot): [0, T] \rightarrow \mathcal{L}(Z, X)$ is said to be uniformly Borel-measurable if there exists a sequence of Borel-measurable simple functions $F_n(\cdot): [0, T] \rightarrow \mathcal{L}(X, Z)$ such that

$$(A.11) \quad \lim_{n \rightarrow \infty} \|F_n(t) - F(t)\|_{\mathcal{L}(Z, X)} = 0 \quad \forall t \in [0, T].$$

It is easy for us to define the Bochner integral for the operator-valued function $F(\cdot):[0, T] \rightarrow \mathcal{L}(Z, X)$ with respect to the vector measure $\bar{\mu}$ induced by $\mu \in BV_0([0, T]; Z)$. We denote

$$B_\mu([0, T]; \mathcal{L}(Z, X)) = \{F:[0, T] \rightarrow \mathcal{L}(Z, X) \mid F(\cdot) \text{ is } \bar{\mu}\text{-Bochner-integrable}\}.$$

Next, we would like to introduce the Young integral.

DEFINITION A.2. Let $\mu(\cdot) \in BV_0([0, T]; K)$ and $F(\cdot):[0, T] \rightarrow \mathcal{L}(Z, X)$ be uniformly Borel-measurable. We say that $F(\cdot)$ is μ -Young-integrable if $F(\cdot)$ is $\bar{\mu}_c$ -Bochner-integrable and

$$(A.12) \quad \sum_{0 \leq \tau \leq T} \|F(\tau)\|_{\mathcal{L}(Z, X)} |\mu(\tau+0) - \mu(\tau-0)| < \infty.$$

In this case, we define the Young integral of $F(\cdot)$ with respect to $\mu(\cdot)$ by the following:

$$(A.13) \quad \begin{aligned} \int_s^t F(\tau) d\mu(\tau) &= \int_{[s, t]} F(\tau) \bar{\mu}_c(d\tau) + \sum_{s < \tau < t} F(\tau) [\mu(\tau+0) - \mu(\tau-0)] \\ &\quad + F(s) [\mu(s+0) - \mu(s)] + F(t) [\mu(t) - \mu(t-0)], \\ &\quad \forall 0 \leq s < t \leq T, \end{aligned}$$

$$\int_t^t F(\tau) d\mu(\tau) = 0 \quad \forall t \in [0, T].$$

We let

$$Y_\mu([0, T]; \mathcal{L}(Z, X)) = \{F:[0, T] \rightarrow \mathcal{L}(Z, X) \mid F(\cdot) \text{ is } \mu\text{-Young-integrable}\}.$$

Then we have the following result concerning the relation between Bochner and Young integrals.

PROPOSITION A.3. Let $\mu(\cdot) \in BV_0([0, T]; K)$. Then

$$(A.14) \quad Y_\mu([0, T]; \mathcal{L}(Z, X)) = B_\mu([0, T]; \mathcal{L}(Z, X)),$$

and, for any $F(\cdot) \in B_\mu([0, T]; \mathcal{L}(Z, X))$, we have that

$$(A.15) \quad \begin{aligned} \int_s^t F(\tau) d\mu(\tau) &= \int_{[s, t]} F(\tau) \bar{\mu}(d\tau) - F(s) [\mu(s) - \mu(s-0)] \\ &\quad - F(t) [\mu(t+0) - \mu(t)] \\ &= \int_{(s, t)} F(\tau) \bar{\mu}(d\tau) + F(s) [\mu(s+0) \\ &\quad - \mu(s)] + F(t) [\mu(t) - \mu(t-0)], \\ &\quad \forall 0 \leq s < t \leq T. \end{aligned}$$

In particular,

$$(A.16) \quad \int_0^T F(\tau) d\mu(\tau) = \int_{[0, T]} F(\tau) \bar{\mu}(d\tau).$$

A similar result for the case where $F(\cdot)$ and $\mu(\cdot)$ are scalar-valued can be found in [18]. Since, in our optimal control problem, the vector measure $\bar{\mu}(\cdot)$ (or the BV function $\mu(\cdot)$) will be a part of control, we need the following notion.

DEFINITION A.4. A function $F(\cdot):[0, T] \rightarrow \mathcal{L}(Z, X)$ is said to be $BV_0([0, T]; K)$ -integrable if for any $\mu(\cdot) \in BV_0([0, T]; K)$, $F(\cdot)$ is $\bar{\mu}$ -Bochner-integrable, or, equivalently, μ -Young-integrable.

PROPOSITION A.5. Let $F(\cdot):[0, T] \rightarrow \mathcal{L}(Z, X)$ be uniformly Borel-measurable. Then $F(\cdot)$ is $BV_0([0, T]; Z)$ -integrable if and only if

$$(A.17) \quad \|F(\tau)\|_{\mathcal{L}(Z, X)} \leq C, \quad \tau \in [0, T].$$

Proof. The sufficiency is obvious. To prove the necessity, we suppose the contrary. Then there exist sequences $\tau_i \in [0, T]$ and $z_i \in Z$ with $|z_i| = 1/i^2$ ($i \geq 1$) such that

$$(A.18) \quad |F(\tau_i)z_i| \geq 1 \quad \forall i \geq 1.$$

Then let

$$\mu(t) = \sum_{i \geq 1} z_i \chi_{[\tau_i, \infty)}(t), \quad t \in [0, T].$$

We see that this $\mu(\cdot) \in BV_0([0, T]; Z)$ and $F(\cdot)$ is not $\bar{\mu}$ -Bochner-integrable. \square

Next, we introduce the μ -continuity. Let $\mu(\cdot) \in BV_0([0, T]; K)$. We introduce a metric on $[0, T]$, which is related to $\mu(\cdot)$ and is stronger than the usual Euclidean metric, as follows:

$$(A.19) \quad \rho_\mu(s, t) = |s - t| + \|\mu|(s) - \mu|(t)\| \quad \forall s, t \in [0, T].$$

Then, for any metric space V , we define

$$(A.20) \quad C_\mu([0, T]; V) = \{v(\cdot): [0, T] \rightarrow V \mid v(\cdot) \text{ is uniformly continuous from } [0, T] \text{ with metric } \rho_\mu(\cdot) \text{ to } V\}.$$

We should note that, in general, $[0, T]$ is not necessarily compact under metric $\rho_\mu(\cdot)$ (see [18] for relevant remarks on the scalar-valued case). Thus, in (A.20), we require $v(\cdot)$ to be uniformly continuous, instead of just continuous.

PROPOSITION A.6. Let $\mu \in BV_0([0, T]; K)$. Then

$$(A.21) \quad \mu(\cdot) \in C_\mu([0, T]; K),$$

and, if $v(\cdot) \in C_\mu([0, T]; V)$, then

$$(A.22) \quad \{t \in [0, T] \mid v(t-0) \neq v(t)\} \subseteq \{t \in [0, T] \mid \mu(t-0) \neq \mu(t)\},$$

$$(A.23) \quad \{t \in [0, T] \mid v(t+0) \neq v(t)\} \subseteq \{t \in [0, T] \mid \mu(t+0) \neq \mu(t)\}.$$

In particular,

$$(A.24) \quad C([0, T]; V) \subseteq C_\mu([0, T]; V) \quad \forall \mu(\cdot) \in BV_0([0, T]; K),$$

$$(A.25) \quad C([0, T]; V) = C_\mu([0, T]; V) \quad \forall \mu(\cdot) \in BV_0([0, T]; K) \cap C([0, T]; K).$$

Proof. By (A.19), we see that

$$(A.26) \quad |\mu(t) - \mu(s)| \leq \rho_\mu(t, s) \quad \forall t, s \in [0, T].$$

Thus (A.21) follows. Now we let $v(\cdot) \in C_\mu([0, T]; V)$ and d be the metric of V . Then there exists a modulus of continuity $\omega(\cdot)$ such that

$$(A.27) \quad d(v(s), v(t)) \leq \omega(\rho_\mu(s, t)) \quad \forall s, t \in [0, T].$$

Thus, if $t \in (0, T]$ with the property $\mu(t-0) = \mu(t)$, then

$$(A.28) \quad \begin{aligned} d(v(t), v(t-\varepsilon)) &\leq \omega(\rho_\mu(t, t-\varepsilon)) \\ &\leq \omega(\varepsilon + |\mu|(t) - |\mu|(t-\varepsilon)) \rightarrow 0 \quad (\varepsilon \rightarrow 0). \end{aligned}$$

Thus (A.22) follows. Similarly, we can obtain (A.23). Then (A.24) follows easily. To obtain (A.25), we let $\mu(\cdot) \in BV_0([0, T]; K) \cap C([0, T]; K)$. Then $|\mu|(\cdot) \in C([0, T]; \mathbb{R})$. Thus, for some modulus of continuity $\omega(\cdot)$, we have that

$$(A.29) \quad ||\mu|(s) - |\mu|(t)| \leq \omega(|s - t|) \quad \forall s, t \in [0, T].$$

Hence, if $v(\cdot) \in C_\mu([0, T]; V)$, then, for some modulus of continuity $\hat{\omega}(\cdot)$, we have that

$$(A.30) \quad d(v(s), v(t)) \leq \hat{\omega}(\rho_\mu(s, t)) \leq \hat{\omega}(|s - t| + \omega(|s - t|)).$$

Hence (A.25) follows. \square

Any function $v(\cdot) \in C_\mu([0, T]; V)$ is said to be μ -continuous. The main features of the μ -continuous functions are that they at most jump at the points where $\mu(\cdot)$ jumps, and at every point $t \in [0, T]$, the left and the right limits exist. We should note, however, that these functions are not necessarily of bounded variation.

Next, we consider the indefinite integral of the Young type. To this end, let $\mu(\cdot) \in BV_0([0, T]; K)$ and $F(\cdot) \in Y_\mu([0, T]; \mathcal{L}(Z, X))$. Then, for any $t \in [0, T]$, we have that

$$(A.31) \quad \int_0^t F(\tau) d\mu(\tau) = \int_{[0, t]} F(\tau) \bar{\mu}_c(d\tau) + \sum_{0 < \tau < t} F(\tau) [\mu(\tau + 0) - \mu(\tau - 0)] \\ + F(0) [\mu(0 + 0) - \mu(0)] + F(t) [\mu(t) - \mu(t - 0)].$$

We refer to the above indefinite Young integral.

PROPOSITION A.7. Let $\mu(\cdot) \in BV_0([0, T]; K)$ and $F(\cdot) \in Y_\mu([0, T]; \mathcal{L}(Z, X))$. Let

$$(A.32) \quad f(t) = \int_0^t F(\tau) d\mu(\tau) \quad \forall t \in [0, T].$$

Then $f(\cdot) \in C_\mu([0, T]; X) \cap BV([0, T]; X)$, and

$$(A.33) \quad f(t + 0) = f(t) + F(t) [\mu(t + 0) - \mu(t)], \quad t \in [0, T],$$

$$(A.34) \quad f(t - 0) = f(t) - F(t) [\mu(t) - \mu(t - 0)], \quad t \in (0, T],$$

$$(A.35) \quad |f|(t) \leq \int_0^t \|F(\tau)\|_{\mathcal{L}(Z, X)} d|\mu|(\tau), \quad t \in [0, T].$$

We let $\Delta = \{(t, \tau) \in [0, T] \times [0, T] \mid t \geq \tau\}$ and $\bar{\Delta}$ be its closure. Let $G(\cdot, \cdot): \bar{\Delta} \rightarrow \mathcal{L}(Z, X)$, and consider the following integral:

$$(A.36) \quad g(t) = \int_0^t G(t, \tau) d\mu(\tau), \quad t \in [0, T].$$

We have the following result, the proof of which is straightforward.

PROPOSITION A.8. Let $\mu(\cdot) \in BV_0([0, T]; K)$ and $G(\cdot, \cdot): \bar{\Delta} \rightarrow \mathcal{L}(Z, X)$. We assume the following:

(i) $G(\cdot, \cdot)$ is uniformly Borel-measurable. There exists a $|\bar{\mu}|$ -integrable function $a(\cdot): [0, T] \rightarrow \mathbb{R}$ such that

$$(A.37) \quad \|G(t, \tau)\|_{\mathcal{L}(Z, X)} \leq a(\tau) \quad \forall (t, \tau) \in \bar{\Delta}.$$

(ii) There exist functions $\omega_G: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and $\theta: [0, T] \rightarrow \mathbb{R}^+$, with the properties that ω_G is nondecreasing, $\omega_G(0) = 0$, and $\theta(\cdot)$ is $|\bar{\mu}|$ -integrable, such that

$$(A.38) \quad \|G(t, \tau) - G(s, \tau)\|_{\mathcal{L}(Z, X)} \leq \omega_G(\rho_\mu(t, s))\theta(\tau) \quad \forall (t, \tau), (s, \tau) \in \bar{\Delta}.$$

Then the function $g(\cdot)$ defined by (A.36) is μ -continuous and

$$(A.39) \quad \begin{aligned} g(t+0) &= \int_0^t G(t+0, \tau) d\mu(\tau) + G(t+0, t)[\mu(t+0) - \mu(t)], \\ g(t-0) &= \int_0^t G(t-0, \tau) d\mu(\tau) - G(t-0, t)[\mu(t) - \mu(t-0)], \end{aligned} \quad t \in [0, T],$$

where

$$(A.40) \quad G(t-0, t) \equiv G(t, t), \quad t \in [0, T], \quad G(T+0, T) = G(T, T).$$

In particular, if there exists a constant C such that

$$(A.41) \quad \omega_G(r) = Cr \quad \forall r \in \mathbb{R},$$

then $g(\cdot) \in BV_0([0, T]; X)$.

Next, we consider a Young-type multiple integral and the corresponding Fubini theorem. We let X , Y , and Z be a Banach space and let $T, S > 0$ be constants. Let $\mu(\cdot) \in BV_0([0, T]; Z)$ and $\nu(\cdot) \in BV_0([0, S]; Y)$. We let

$$(A.42) \quad \begin{aligned} \tilde{\mathcal{B}}(Y \times Z; X) &= \left\{ B: Y \times Z \rightarrow X \mid B \text{ is bilinear and} \right. \\ &\quad \left. \|B\| \equiv \sup_{|y|_X, |z|_Z \leq 1} |B(y, z)|_X < \infty \right\}. \end{aligned}$$

It is easy to see that $\tilde{\mathcal{B}}(Y \times Z; X)$ is a Banach space. Thus we may define the uniform Borel measurability for $\tilde{\mathcal{B}}(Y \times Z; X)$ -valued functions and the multiple Bochner-type integral of such functions with respect the vector measures $\bar{\mu}$ and $\bar{\nu}$. Then we can introduce the $(\mu \times \nu)$ -Young integral, which is defined in the following way: Let $\Psi: [0, T] \times [0, S] \rightarrow \tilde{\mathcal{B}}(Y \times Z; X)$ be uniformly Borel-measurable and

$$(A.43) \quad \int_{[0, T] \times [0, S]} \|\Psi(t, s)\| |\bar{\mu}|(dt) |\bar{\nu}|(ds) < \infty.$$

Then, for $0 \leq t_0 \leq t_1 \leq T$ and $0 \leq s_0 \leq s_1 \leq S$, we define

$$(A.44) \quad \begin{aligned} \int_{t_0}^{t_1} \int_{s_0}^{s_1} \Psi(t, s, d\mu(t), d\nu(s)) &= \int_{t_0}^{t_1} \int_{[s_0, s_1]} \Psi(t, s, d\mu(t), \bar{\nu}(ds)) \\ &\quad - \int_{t_0}^{t_1} \Psi(t, s_0, d\mu(t), \nu(s_0) - \nu(s_0 - 0)) \\ &\quad - \int_{t_0}^{t_1} \Psi(t, s_1, d\mu(t), \nu(s_1 + 0) - \nu(s_1)). \end{aligned}$$

The meaning of the right-hand side of (A.44) is clear ($\int d\mu(t)$ always means the Young integral, and $\int \bar{\nu}(ds)$ means the Bochner integral). In the case where the above Young integral can be defined, we say that the function Ψ is $(\mu \times \nu)$ -Young-integrable. From our definition, the $(\bar{\mu} \times \bar{\nu})$ -Bochner integrability and the $(\mu \times \nu)$ -Young integrability are the same. Thus, hereafter, we simply refer to these two integrabilities as the $(\mu \times \nu)$ -integrability. Next, let us introduce the following notation (compare (A.13)):

$$(A.45) \quad \begin{aligned} \int_s^t F(\tau) d\mu(\tau) &= \int_{[s, t]} F(\tau) \bar{\mu}_c(d\tau) + \sum_{s < \tau < t} F(\tau)[\mu(\tau+0) - \mu(\tau-0)] \\ &\quad + F(t)[\mu(t) - \mu(t-0)], \end{aligned}$$

$$(A.46) \quad \int_s^t F(\tau) d\mu(\tau) = \int_{[s,t]} F(\tau) \bar{\mu}_c(d\tau) + \sum_{s \leq \tau < t} F(\tau) [\mu(\tau+0) - \mu(\tau-0)] \\ + F(t) [\mu(t) - \mu(t-0)],$$

$$(A.47) \quad \int_s^{(t)} F(\tau) d\mu(\tau) = \int_{[s,t]} F(\tau) \bar{\mu}_c(d\tau) + \sum_{s < \tau < t} F(\tau) [\mu(\tau+0) - \mu(\tau-0)] \\ + F(s) [\mu(s+0) - \mu(s)],$$

$$(A.48) \quad \int_s^{[t]} F(\tau) d\mu(\tau) = \int_{[s,t]} F(\tau) \bar{\mu}_c(d\tau) + \sum_{s < \tau \leq t} F(\tau) [\mu(\tau+0) - \mu(\tau-0)] \\ + F(s) [\mu(s+0) - \mu(s)].$$

Similarly, we can define $\int_s^{(t)} F(\tau) d\mu(\tau)$, $\int_{[s,t]} F(\tau) d\mu(\tau)$, $\int_s^{[t]} F(\tau) d\mu(\tau)$ and $\int_{[s,t]} F(\tau) d\mu(\tau)$. It is easy to see that

$$(A.49) \quad \int_{(s)}^{(t)} F(\tau) d\mu(\tau) = \int_{(s,t)} F(\tau) \bar{\mu}(d\tau), \quad \int_{[s]}^{(t)} F(\tau) d\mu(\tau) = \int_{[s,t)} F(\tau) \bar{\mu}(d\tau), \\ \int_{(s)}^{[t]} F(\tau) d\mu(\tau) = \int_{(s,t]} F(\tau) \bar{\mu}(d\tau), \quad \int_{[s]}^{[t]} F(\tau) d\mu(\tau) = \int_{[s,t]} F(\tau) \bar{\mu}(d\tau).$$

The following result is a kind of Fubini theorem for the $(\mu \times \nu)$ -Young integral. The result has its own interest. To our knowledge, even for the scalar case of the Young integral, such a result is new. In deriving the adjoint system along the optimal triplet in § 5, it played an important role.

THEOREM A.9 (the Fubini theorem). *Let $\mu(\cdot) \in BV([0, T]; Z)$, $\nu(\cdot) \in BV([0, T]; Y)$ and $B(\cdot, \cdot): [0, T] \times [0, S] \rightarrow \mathcal{B}(Y \times Z; X)$ be $(\mu \times \nu)$ -integrable. Then*

$$(A.50) \quad \int_s^t \int_\tau^t B(r, \tau) d\mu(r) d\nu(\tau) - \int_s^t \int_s^\tau B(r, \tau) d\nu(\tau) d\mu(r) \\ = \int_s^{(t)} B(\tau, \tau) [\mu(\tau+0) - \mu(\tau)] d\nu(\tau) \\ - \int_{(s)}^t B(r, r) [\nu(r) - \nu(r-0)] d\mu(r)$$

$$\forall 0 \leq s < \tau < t \leq T.$$

In particular, if

$$(A.51) \quad B(\tau, \tau) = 0 \quad \forall 0 \leq \tau \leq T,$$

then

$$(A.52) \quad \int_s^t \int_\tau^t B(r, \tau) d\mu(r) d\nu(\tau) = \int_s^t \int_s^\tau B(r, \tau) d\nu(\tau) d\mu(r).$$

In the above, we have taken the convention that

$$B(t, s)yz = B(t, s)zy \quad \forall (y, z) \in Y \times Z.$$

This will not cause any confusion from the context. The proof of the above result consists of some direct computations. We see that in the case where (A.51) fails, the

right-hand side of (A.50) is not necessarily zero. Thus, when changing the order of Young integrals, we must be somewhat careful.

REFERENCES

- [1] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, North-Holland, New York, 1981.
- [2] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
- [3] H. T. BANKS AND G. KENT, *Control of functional equations to target sets in function space*, SIAM J. Control Optim., 10 (1972), pp. 567–593.
- [4] V. BARBU, *Optimal Control of Variational Inequalities*, Pitman, Boston, 1984.
- [5] G. BARLES, *Deterministic impulse control problems*, SIAM J. Control Optim., 23 (1985), pp. 419–432.
- [6] S. A. BELBAS AND S. M. LENHART, *Nonlinear PDE's for stochastic optimal control with switching and impulses*, Appl. Math. Optim., 14 (1986), pp. 215–227.
- [7] A. BENSOUSSAN AND J. L. LIONS, *Impulse Control and Quasi-Variational Inequalities*, Bordes, Paris, 1984.
- [8] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
- [9] F. H. CLARKE, *The maximum principle with minimum hypotheses*, SIAM J. Control Optim., 14 (1976), pp. 1078–1091.
- [10] ———, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [11] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, Vol. 8, Springer-Verlag, New York, 1981.
- [12] P. C. DAS AND R. R. SHARMA, *On optimal controls for measure delay-differential equations*, SIAM J. Control Optim., 9 (1971), pp. 43–61.
- [13] J. DIESTEL AND J. J. UHL, JR., *Vector Measures*, American Mathematical Society, Providence, RI, 1977.
- [14] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Interscience, London, 1958.
- [15] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.
- [16] ———, *Nonconvex minimization problems*, Bull. Amer. Math. Soc., 1 (1979), pp. 443–474.
- [17] H. O. FATTORINI, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Appl. Math. Optim., 15 (1987), pp. 141–185.
- [18] J. GROH, *A nonlinear Volterra–Stieltjes integral equation and a Gronwall inequality in one dimension*, Illinois J. Math., 34 (1980), pp. 244–263.
- [19] T. H. HILDEBRANDT, *On systems of linear differentio-Stieltjes-integral equations*, Illinois J. Math., 3 (1959), pp. 352–373.
- [20] E. HILLE AND R. S. PHILIPS, *Functional Analysis and Semigroups*, American Mathematical Society, Providence, RI, 1957.
- [21] C. S. HÖNIG, *Volterra Stieltjes Integral Equations*, North-Holland, Amsterdam, 1975.
- [22] X. LI AND Y. YAO, *Maximum principle of distributed parameter systems with time lags*, in *Distributed Parameter Systems*, Lecture Notes in Control and Information Sciences, Vol. 75, Springer-Verlag, New York, 1985, pp. 410–427.
- [23] X. LI AND J. YONG, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.
- [24] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [25] J. L. MENALDI, *Optimal impulse control problems for degenerate diffusions with jumps*, Acta Appl. Math., 8 (1987), pp. 165–198.
- [26] A. B. MINGARELLI, *Volterra–Stieltjes Integral Equations and Generalized Ordinary Differential Expressions*, Lecture Notes in Mathematics 989, Springer-Verlag, New York, 1983.
- [27] I. P. NATANSON, *Theory of Functions of a Real Variable*, Ungar, New York, 1955.
- [28] S. G. PANDIT AND S. G. DEO, *Differential Systems Involving Impulses*, Lecture Notes in Mathematics 954, Springer-Verlag, New York, 1982.
- [29] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [30] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
- [31] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control law contain measures*, SIAM J. Control Optim., 3 (1965), pp. 191–205.

- [32] R. B. VINTER AND F. M. F. L. PEREIRA, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205–229.
- [33] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [34] J. YONG, *Optimal switching and impulse controls for distributed parameter systems*, System Sci. Math. Sci., 2 (1989), pp. 137–160.
- [35] ———, *Systems governed by ordinary differential equations with continuous, switching and impulse controls*, Appl. Math. Optim., 20 (1989), pp. 223–236.
- [36] ———, *Maximum principle of optimal controls for a nonsmooth semilinear evolution system*, in Analysis and Optimization of Systems, A. Benssoussan and J. L. Lions, eds., Lecture Notes in Control Inform. Sci., Springer-Verlag, 144 (1990), pp. 559–569.
- [37] J. YONG AND P. ZHANG, *Necessary conditions of optimal impulse controls for distributed parameter systems*, Bull. Austral. Math. Sci., 45 (1992), pp. 305–326.
- [38] W. H. YOUNG, *On integration with respect to a function of bounded variation*, Proc. London Math. Soc. (2), 13 (1914), pp. 109–150.

LIPSCHITZIAN STABILITY IN NONLINEAR CONTROL AND OPTIMIZATION*

ASEN L. DONTCHEV† AND WILLIAM W. HAGER‡

Abstract. This paper studies Lipschitz properties, relative to the parameter p , of the set of solutions to problems of the form

$$\text{Find } z \in \Omega_p \text{ such that } T_p(z) \in F_p(z).$$

As applications, various problems in control and optimization are examined, focusing in particular on the stability of the feasible set of a control problem, and the stability of solutions of infinite-dimensional mathematical programs and optimal control problems. In another application, an estimate is obtained for the error in the Euler approximation to an optimal control problem.

Key words. stability, sensitivity, feasible set, optimal set, discrete approximations, Euler approximation, sufficient optimality conditions

AMS(MOS) subject classifications. 49K40, 49M25, 90C31, 93B05

1. Introduction. This paper presents a general framework for analyzing Lipschitz stability in control and optimization. As applications of the theory, we study the dependence on a parameter of the set of controls and states that satisfies given inequality constraints. We also study the dependence on a parameter of the optimal solutions of various problems in nonlinear control and optimization.

The paper begins by studying (in § 2) the following problem:

$$(1) \quad \text{Find } z \in \Omega_p \text{ such that } T_p(z) \in F_p(z),$$

where p is a parameter, T_p maps Ω_p to Y_p , Y_p is a normed linear space, and $F_p(z)$ is a subset of Y_p for each $z \in \Omega_p$. Loosely speaking, we proceed in the following way: Along with (1), we consider an *auxiliary problem*

$$(2) \quad \text{Find } z \in \Omega_p \text{ such that } L_p(z) + y \in F_p(z),$$

where $y \in Y$ is treated as a new parameter. It turns out that if L_p approximates T_p in a suitable sense, and if the set of solutions of (2) possesses certain Lipschitz properties with respect to y , uniformly in p , then the set of solutions of (1) will have analogous properties with respect to p . In particular, if T_p is smooth, then L_p can be its linearization. For a nonsmooth T_p , we should choose a nonsmooth L_p .

Our abstract approach is based on a refinement of the set-valued contracting mapping principle (Lemma 1). An existence result given in Theorem 1, leads to various stability results. In particular, Corollary 1 obtains an estimate for the distance from a reference point to the set of solutions of (1). In Corollary 2, we assume that Ω , F , and L are independent of p , obtaining an implicit function theorem: If the solution set of (2) is pseudo-Lipschitz with respect to y around some given point, and if L strongly approximates T_p , then the set of solutions of (1) is pseudo-Lipschitz as well. Corollary 3 obtains a result related to metric regularity of the map $T - F$.

* Received by the editors July 9, 1990; accepted for publication (in revised form) December 4, 1991. This research was supported by the United States Army Research Office contract DAAL03-89-G-0082, by National Science Foundation grant DMS 9022899, and by the Bulgarian Ministry of Science contract 127. The work was performed while the first author was a visitor at the University of Florida.

† Mathematical Reviews, 416 Fourth Street, Ann Arbor, Michigan 48107 (on leave from the Institute of Mathematics, Bulgarian Academy of Sciences, Sofia, Bulgaria).

‡ Department of Mathematics, University of Florida, Gainesville, Florida 32611.

Generalized equations of the form (1) have been considered by Robinson in a series of papers [35]–[38] with Ω and F independent of p . Our analysis contains some of his results. While the theory of [35]–[38] is applied to finite-dimensional mathematical programming problems, our focus here is infinite-dimensional optimization, primarily optimal control. Our analysis is more in the spirit of [19] and [20].

Recently, a different approach to sensitivity based on nonsmooth analysis and the differentiability properties of set-valued maps was developed by Aubin [4], Aubin and Frankowska [6], Rockafellar [39] and [41], King and Rockafellar [24], Mordukhovich [32], and others. In [16] this approach is applied to various control problems. A motivation for the nonsmooth approach to Lipschitz stability is given by Rockafellar in [40].

An outline of our paper follows, while detailed comments connecting specific results in our paper to related literature appear throughout the paper. Section 3 examines the feasible set for a nonlinear control system with inequality state and control constraints that depend on a parameter. We show that if the functions defining the constraints are sufficiently smooth, and if an interior point condition holds for a linearized system, then the feasible set is pseudo-Lipschitz. Moreover, the interior point condition holds if the gradients of the active constraints satisfy an independence condition, the same condition that appeared in Hager's analysis [18] of Lipschitz continuity in time for an optimal control. At the end of § 3, we present an example of a nonsmooth control system with state and control constraints, and we demonstrate a method for proving local controllability.

Section 4 considers a quadratic minimum problem in a reflexive Banach space with linear cone constraints. We show that a coercivity condition together with surjectivity of the gradients of the (active) constraints guarantee local Lipschitz continuity of the solution relative to the data. In § 5 we apply this result to a nonlinear optimization problem, and a quadratic program plays the role of an auxiliary problem.

In § 6 we consider a nonlinear control problem with convex control constraints. The treatment of the control problem requires special care due to the discrepancy between the function spaces needed for coercivity and for differentiability. An example shows that the method of analysis can still be applied, even when the coercivity condition is violated.

Finally, in § 7 we obtain error estimates for Euler's approximation to a nonlinear optimal control problem with convex control constraints. In this case, the parameter p in (1) corresponds to the mesh spacing. The key step in the analysis is to show that the solution of a perturbed discrete linear-quadratic problem, related to the auxiliary problem (2), depends Lipschitz continuously on a parameter, uniformly in the mesh spacing. Our method makes use of the so-called averaged modulus of smoothness, introduced by Sendov and Popov [42]. When the optimal control has bounded variation, the error in the discrete control is on the order of the mesh spacing.

2. Abstract theory. Let Z be a Banach space, let Ω be a closed subset of Z , let Y be a normed vector space, and let 2^Y denote the collection of subsets of Y . Given a map $T: \Omega \rightarrow Y$ and a map $F: \Omega \rightarrow 2^Y$, we consider the following problem:

$$(3) \quad \text{Find } z \in \Omega \text{ such that } T(z) \in F(z).$$

Of course, for appropriate choices of Ω , T , and F , (3) may represent an equation, an inclusion, or a variational inequality. In this section, conditions are formulated that guarantee a solution to (3). This existence theorem applied to perturbations of (3) yields stability results. Throughout this paper, $\|\cdot\|$ denotes a norm in the appropriate

space. Given subsets P and $Q \subset Z$, the one-sided distance from P to Q (or excess function), denoted $\|P - Q\|$, is defined by

$$\|P - Q\| = \sup_{p \in P} \inf_{q \in Q} \|p - q\|.$$

If Q is empty, we set $\|P - Q\| = \infty$. Given $z \in Z$, let $B_r(z)$ denote the closed ball with center z and radius r .

We will use a contraction mapping principle for set-valued maps to obtain an existence result for (3). The proof that follows is similar to the usual proofs for the existence of a fixed point (see [22, p. 31] or [34]); however, since our Lipschitz assumption (b) below is weaker than the usual Lipschitz assumption, we include a proof for completeness. Although this fixed point result is stated for a Banach space, it holds in any complete metric space.

LEMMA 1. Let $\Phi: \Omega \rightarrow 2^\Omega$ with $\Phi(z)$ closed for every $z \in \Omega$. Suppose that there exist real numbers r and λ , and $z_0 \in \Omega$ with the following properties:

$$(a) \quad 0 \leq \lambda < 1 \quad \text{and} \quad \frac{\|z_0 - \Phi(z_0)\|}{1 - \lambda} < r,$$

$$(b) \quad \|\Phi(y) \cap B_r(z_0) - \Phi(z)\| \leq \lambda \|y - z\| \text{ for every } y \text{ and } z \in B_r(z_0) \cap \Omega.$$

Then there exists $z \in B_r(z_0) \cap \Omega$ such that $z \in \Phi(z)$. If Φ is single-valued, then assumption (a) can be replaced by

$$(a') \quad 0 \leq \lambda < 1 \quad \text{and} \quad \frac{\|z_0 - \Phi(z_0)\|}{1 - \lambda} \leq r,$$

and there exists a unique $z \in B_r(z_0) \cap \Omega$ with $z = \Phi(z)$.

Proof. By assumption (a), there exists $z_1 \in \Phi(z_0)$ such that $\|z_1 - z_0\| < r(1 - \lambda)$. Proceeding by induction, suppose that there exists $z_{k+1} \in \Phi(z_k) \cap B_r(z_0)$ for $k = 1, 2, \dots, n-1$ with $\|z_{k+1} - z_k\| < r(1 - \lambda)\lambda^k$. By assumption (b) and the induction hypothesis, we have

$$\|z_n - \Phi(z_n)\| \leq \|\Phi(z_{n-1}) \cap B_r(z_0) - \Phi(z_n)\| \leq \lambda \|z_n - z_{n-1}\| < r(1 - \lambda)\lambda^n.$$

Hence, there exists $z_{n+1} \in \Phi(z_n)$ such that $\|z_{n+1} - z_n\| < r(1 - \lambda)\lambda^n$. By the triangle inequality,

$$\|z_{n+1} - z_0\| \leq \sum_{k=0}^n \|z_{k+1} - z_k\| < r(1 - \lambda) \sum_{k=0}^n \lambda^k < r,$$

so that $z_{n+1} \in B_r(z_0)$. This completes the induction step.

By the triangle inequality and for $n > m$, we have

$$\|z_n - z_m\| \leq \sum_{k=m}^{n-1} \|z_{k+1} - z_k\| \leq r(1 - \lambda) \sum_{k=m}^{n-1} \lambda^k < r\lambda^m.$$

Thus the z_k form a Cauchy sequence that converges to some limit $z \in B_r(z_0) \cap \Omega$. By assumption (b),

$$\|z_k - \Phi(z)\| \leq \|\Phi(z_{k-1}) \cap B_r(z_0) - \Phi(z)\| \leq \lambda \|z_{k-1} - z\|.$$

Again, by the triangle inequality,

$$\|z - \Phi(z)\| \leq \|z - z_k\| + \|z_k - \Phi(z)\| \leq \|z - z_k\| + \lambda \|z_{k-1} - z\|,$$

which approaches zero as k increases. Since $\Phi(z)$ is closed, it follows that $z \in \Phi(z)$. If Φ is single-valued, then by (b), z is the unique element in $B_r(z_0) \cap \Omega$ for which $z = \Phi(z)$. \square

Given $y \in Y$ and a mapping L from Z to Y , consider the *auxiliary problem*

$$(4) \quad \text{Find } z \in \Omega \text{ such that } L(z) + y \in F(z).$$

Note that the set of solutions to (4) is closed if L is continuous and the graph of F is closed. Given $z_0 \in Z$ and $y_0 \in Y$, define the parameters D_r and δ by

$$D_r = \sup_{\substack{y, z \in B_r(z_0) \cap \Omega \\ y \neq z}} \frac{\|T(z) - T(y) - L(z) + L(y)\|}{\|z - y\|} \quad \text{and} \quad \delta = \|T(z_0) - L(z_0) - y_0\|.$$

If L is a bounded, linear operator, and z_0 lies in the interior of Ω , then $D_r \rightarrow 0$ when $r \rightarrow 0$ if and only if T is strictly (Fréchet) differentiable at z_0 (see [5, p. 16]). We have the following generalization of [20, Thm. 1].

THEOREM 1. *Let γ and r be real numbers that satisfy the relations*

$$(5) \quad 0 \leq \gamma D_r < 1 \quad \text{and} \quad r > \frac{\gamma \delta}{1 - \gamma D_r}.$$

Defining the set

$$\Delta_r = \bigcup_{z \in B_r(z_0) \cap \Omega} \{T(z) - L(z)\},$$

let Ψ denote a map from Δ_r to 2^Ω with the following properties: $z_0 \in \Psi(y_0)$, $\Psi(y)$ is a closed, nonempty subset of the solutions to (4) for each $y \in \Delta_r$, and

$$(6) \quad \|\Psi(y_1) \cap B_r(z_0) - \Psi(y_2)\| \leq \gamma \|y_1 - y_2\| \text{ for every } y_1 \text{ and } y_2 \in \Delta_r \cup \{y_0\}.$$

Then (3) has a solution $z \in B_r(z_0)$. If in addition there is only one solution of (4) for every $y \in \Delta_r$, then z is the unique solution of (3) in $B_r(z_0)$, and the second condition in (5) can be weakened to $r \geq \gamma \delta / (1 - \gamma D_r)$.

Proof. We apply Lemma 1 with $\Phi(z) = \Psi(T(z) - L(z))$. Thus z is a solution to (3) if $z \in \Phi(z)$. By (6), we have

$$\|\Phi(z) \cap B_r(z_0) - \Phi(y)\| \leq \gamma \|T(z) - T(y) - L(z) + L(y)\| \leq \gamma D_r \|z - y\|$$

whenever y and $z \in B_r(z_0) \cap \Omega$. Hence, Φ satisfies (b) of Lemma 1 with constant $\lambda = \gamma D_r$. Since $z_0 \in \Psi(y_0)$, it follows that

$$\begin{aligned} \|z_0 - \Phi(z_0)\| &\leq \|\Psi(y_0) \cap B_r(z_0) - \Psi(T(z_0) - L(z_0))\| \\ &\leq \gamma \|y_0 + L(z_0) - T(z_0)\| = \gamma \delta. \end{aligned}$$

Dividing this inequality by $1 - \lambda$, we see that $r > \|z_0 - \Phi(z_0)\| / (1 - \lambda)$. Since the contraction property of Lemma 1 holds on $B_r(z_0) \cap \Omega$, there exists $z \in B_r(z_0)$ with $z \in \Phi(z)$. If the solution of (4) is unique for every $y \in \Delta_r$, then Φ is single-valued on $B_r(z_0) \cap \Omega$. By Lemma 1, there exists a unique $z \in B_r(z_0) \cap \Omega$ with $z = \Phi(z)$. Hence, there is a unique solution to (3) in $B_r(z_0)$. \square

Remark 1. Note that if $\sigma \geq r D_r + \delta$, then $\Delta_r \subset B_\sigma(y_0)$. To prove this, we take the norm of the identity

$$T(z) - L(z) - y_0 = [T(z) - T(z_0) - L(z) + L(z_0)] + [T(z_0) - L(z_0) - y_0],$$

where $z \in B_r(z_0) \cap \Omega$, and we apply the triangle inequality to obtain the relation

$$\|T(z) - L(z) - y_0\| \leq r D_r + \delta \leq \sigma.$$

Now we consider a family of equations, each equation depending on a parameter p contained in a metric space P . Associated with each $p \in P$, there is a closed subset Ω_p of a Banach space Z_p , a normed vector space Y_p , and a pair of maps $T_p: \Omega_p \rightarrow Y_p$ and $F_p: \Omega_p \rightarrow 2^{Y_p}$. Analogous to (3), we study the following problem:

$$(7) \quad \text{Find } z \in \Omega_p \text{ such that } T_p(z) \in F_p(z).$$

Let 0 be a fixed element of P . Using Theorem 1, we will study the continuity of the map $p \rightarrow \Sigma(p)$, where $\Sigma(p)$ is the set of solutions of (7), making use of the following auxiliary problem:

$$(8) \quad \text{Find } z \in \Omega_p \text{ such that } L_p(z) + y \in F_p(z),$$

where $L_p: Z_p \rightarrow Y_p$, and $y \in Y_p$. We give three specific results assuming the maps appearing in (7) and (8) satisfy certain conditions near a reference point. The parameters D_r , Δ_r , and δ of Theorem 1 now depend on p as follows:

$$D_r(p) = \sup_{\substack{y, z \in B_r(z_p) \cap \Omega_p \\ y \neq z}} \frac{\|T_p(z) - T_p(y) - L_p(z) + L_p(y)\|}{\|z - y\|},$$

$$\Delta_r(p) = \Delta_r(p, z_p), \quad \text{where } \Delta_r(p, x) = \bigcup_{z \in B_r(x) \cap \Omega_p} \{T_p(z) - L_p(z)\}, \quad \text{and}$$

$$\delta(p) = \|T_p(z_p) - L_p(z_p) - y_p\|.$$

(Although the norms above may depend on p , this dependence is not indicated explicitly. In § 5 we consider a finite-dimensional discretization of an optimal control problem, in which case the norms depend on the mesh spacing.)

COROLLARY 1. *Let Ψ_p denote a map from a neighborhood of y_p to 2^{Ω_p} with the following properties: $z_p \in \Psi_p(y_p)$, $\Psi_p(y)$ is a closed, nonempty subset of the solutions to (8) for each $y \in \Delta_\sigma(p)$, where $\sigma > 0$, and for some γ and $\alpha > 0$, we have*

$$(9) \quad \|\Psi_p(y_1) \cap B_\alpha(z_p) - \Psi_p(y_2)\| \leq \gamma \|y_1 - y_2\| \text{ for every } y_1 \text{ and } y_2 \in \Delta_\sigma(p) \cup \{y_p\}.$$

If $D_r(p)$ and $\delta(p)$ tend to zero as r and p tend to zero, then for each $\gamma^+ > \gamma$ and for each p sufficiently close to zero, (7) has a solution z such that

$$(10) \quad \|z_p - z\| \leq \gamma^+ \|T_p(z_p) - L_p(z_p) - y_p\|.$$

If there is only one solution to (8) for every $y \in \Delta_\sigma(p)$, then z , satisfying (10), is the unique solution of (7) in a neighborhood of z_p .

Proof. Apply Theorem 1 with $\delta = \delta(p)$ and $r = \gamma^+ \delta(p)$. If $\delta = 0$, then z_p is a solution of (7), and (10) holds with $z = z_p$. If $\delta > 0$, then choose p sufficiently close to 0 that

$$\sigma \geq r, \quad \gamma D_r(p) < 1, \quad \alpha \geq r, \quad \text{and} \quad \gamma^+ > \frac{\gamma}{1 - \gamma D_r(p)}.$$

Hence, Theorem 1 yields (10). \square

Now we wish to start with a given solution z_0 of (7) associated with $p = 0$ and show that for small perturbations in the parameter, we can solve the equation, and in some sense, the solution is "well behaved." In this analysis, we allow T to depend on p , while Ω and F are independent of p . That is, the following problem is considered:

$$(11) \quad \text{Find } z \in \Omega \text{ such that } T_p(z) \in F(z).$$

To study the continuity of the solution map, we work with the fixed auxiliary problem (4) (L is independent of p).

We define an analogue $E_r(p)$ of D_r in which T is replaced by T_p , below:

$$(12) \quad E_r(p) = \sup_{\substack{y, z \in B_r(z_0) \cap \Omega \\ y \neq z}} \frac{\|T_p(z) - T_p(y) - L(z) + L(y)\|}{\|z - y\|}.$$

Following the terminology of Robinson [38], we say that $L(z)$ strongly approximates $T_p(z)$ at $z = z_0$ and $p = 0$ if and only if $E_r(p) \rightarrow 0$ as p and r tend to zero. Note that L does not need to be smooth. For example, if $T_p(z) = f_p(g(z))$, where f_p is Fréchet differentiable and g is Lipschitz, but not necessarily differentiable, then $L(z) = f'_0[g(z_0)]g(z)$ strongly approximates $T_p(z)$ at $z = z_0$ and $p = 0$ under appropriate continuity assumptions (see [38]).

In the following corollary, we take $z_p = z_0$ and $y_p = y_0$, and we replace assumption (9) by pseudo-Lipschitz continuity. Recall (see [4]) that the map Ψ is pseudo-Lipschitz with modulus γ , around a point (y_0, z_0) in the graph of Ψ , if there exist neighborhoods V of y_0 and U of z_0 such that

$$\|\Psi(y_1) \cap U - \Psi(y_2)\| \leq \gamma \|y_1 - y_2\|$$

whenever y_1 and $y_2 \in V$. Letting $\Sigma_r(p) = \Sigma(p) \cap B_r(z_0)$ denote the restriction of $\Sigma(p)$ to $B_r(z_0)$, we have the following corollary.

COROLLARY 2. *Given $z_0 \in \Sigma(0)$, define $y_0 = T_0(z_0) - L(z_0)$, and let $\Psi(y)$ denote the set of solutions to (4). We assume that Ψ is closed and nonempty-valued near y_0 , that Ψ is pseudo-Lipschitz with modulus γ around (y_0, z_0) , and that L strongly approximates $T_p(z)$ at $p = 0$ and $z = z_0$. If $T_p(z)$ is continuous in p at $p = 0$, uniformly in a neighborhood of $z = z_0$, then for each $\gamma^+ > \gamma$ and for r sufficiently small, there exists $s > 0$ such that $\Sigma_r(p)$ is nonempty for every $p \in B_s(0)$; moreover, for each p and $q \in B_s(0)$ and for each $z_p \in \Sigma_r(p)$, there exists $z_q \in \Sigma(q)$ such that*

$$(13) \quad \|z_p - z_q\| \leq \gamma^+ \|T_q(z_p) - T_p(z_p)\|.$$

If there is only one solution to (4) for every y near y_0 , then the solution of (11) is unique in $B_r(z_0)$ for every $p \in B_s(0)$. Moreover, the $z_q \in \Sigma(q)$ satisfying (13) also lies in $B_r(z_0)$.

Proof. Define the parameters

$$d(a, s) = \sup_{\substack{p, q \in B_s(0) \\ z \in B_a(z_0)}} \|T_p(z) - T_q(z)\| \quad \text{and} \quad \bar{D}_\rho = \sup_{p \in B_\rho(0)} E_\rho(p).$$

Let U and V be the neighborhoods of z_0 and y_0 appearing in the definition of pseudo-Lipschitz continuity. Choose σ sufficiently small that $B_\sigma(y_0) \subset V$. Choose ρ sufficiently small that $B_\rho(z_0) \subset U$

$$(14) \quad \sigma > \rho \bar{D}_\rho \quad \text{and} \quad \gamma < \gamma^+(1 - \gamma \bar{D}_\rho).$$

Choose a and s sufficiently small that

$$(15) \quad \rho \geq 2a, \quad \rho \geq s, \quad a > \gamma^+ d(a, s), \quad \text{and} \quad \sigma \geq \rho \bar{D}_\rho + d(a, s).$$

Let $p \in B_s(0)$. Referring to Remark 1, we see that $\Delta_a(p, z_0) \subset B_\sigma(y_0)$. Theorem 1 with T replaced by T_p and with $r = a$ implies that $\Sigma_a(p)$ is nonempty.

Given p and $q \in B_s(0)$ and $z_p \in \Sigma_a(p)$, let us apply Theorem 1 with z_0, y_0 , and T in the theorem replaced by $z_p, y_p = T_p(z_p) - L(z_p)$, and T_q , respectively. In Theorem 1, we take

$$r = \gamma^+ \delta, \quad \text{where} \quad \delta = \|T_q(z_p) - T_p(z_p)\| \leq d(a, s).$$

If $\delta = 0$, then (13) holds trivially by taking $z_q = z_p$. If $\delta > 0$, then by (15) we have

$$r > \frac{\gamma\delta}{1 - \gamma D_\rho},$$

where $D_\rho = E_\rho(q) \leq \bar{D}_\rho$. Since $B_r(z_p) \subset B_\rho(z_0)$, we have $\Delta_r(p, z_p) \subset \Delta_\rho(p, z_0) \subset B_\sigma(y_0)$. Hence, condition (6) of Theorem 1 holds, and (13) is established.

If the map Ψ is single-valued, then there exists a unique solution of (11) in $B_r(z_0)$ for every $p \in B_s(0)$. Similar to the proof of Theorem 1, the map $\Phi_p(z) = \Psi(T_p(z) - L(z))$ is a contraction on the ball $B_a(z_0)$ with contraction constant $\lambda = \gamma\bar{D}_\rho$ for each $p \in B_s(0)$. The distance from z_p to z_q is estimated by the following sequence of inequalities:

$$\begin{aligned} \|z_p - z_q\| &= \|\Phi_p(z_p) - \Phi_q(z_q)\| \leq \|\Phi_p(z_p) - \Phi_q(z_p)\| + \|\Phi_q(z_p) - \Phi_q(z_q)\| \\ &\leq \lambda \|z_p - z_q\| + \gamma \|T_p(z_p) - T_q(z_p)\|, \end{aligned}$$

which yields (13). \square

Observe that if there exists a constant κ such that T_p satisfies the Lipschitz condition

$$\|T_p(z) - T_q(z)\| \leq \kappa \text{ distance } \{p, q\}$$

for every z in a neighborhood of z_0 , then (13) implies that for every $z_p \in \Sigma_a(p)$, there exists $z_q \in \Sigma(q)$ such that

$$\|z_p - z_q\| \leq \gamma^+ \kappa \text{ distance } \{p, q\};$$

that is, the map Σ is pseudo-Lipschitz around $p = 0$ and $z = z_0$. Thus we conclude that if the auxiliary problem strongly approximates the original problem, and if the solution map of the auxiliary problem is pseudo-Lipschitz, then the solution map of the original problem is pseudo-Lipschitz as well.

Remark 2. Corollary 2 is a generalization of Theorem 2.1 in [37] and of Theorem 3.2 in [38]. In [37] Ω is a closed convex set, $F(z)$ is the normal cone to Ω at z , $T_p(z)$ is Fréchet differentiable with respect to z around $z = z_0$ and $p = 0$, and both $T_p(z)$ and its derivative $T'_p(z)$ are continuous with respect to z and p at $z = z_0$ and $p = 0$. Furthermore, it is assumed that (4) with

$$L(z) = T_0(z_0) + T'_0(z_0)(z - z_0)$$

has a unique solution that is Lipschitz near $y_0 = 0$. In [38] $F(z) = 0$, $L(z)$ strongly approximates $T_p(z)$ at $z = z_0$ and $p = 0$, and the assumptions for $L(z)$ are equivalent to the condition that L^{-1} is single-valued and Lipschitz near 0.

Corollary 2 is an implicit function theorem in which we avoid the surjectivity (interiority) condition, for a suitably defined derivative, that is usually present in a Graves-type theorem (see [7, p. 95]). For example, given a closed-valued map $F: Z \rightarrow 2^Y$ and given (z_0, y_0) in the graph of F , let $f: Z \rightarrow Y$ be a continuous function that is strictly Fréchet differentiable at z_0 . Let us apply Corollary 2 with $L(z) = -f'(z_0)(z - z_0)$, $p = y \in Y$, $T_p(z) = p - f(z)$, and Ψ defined by

$$\Psi(y) = \{z \in Z: y - f'(z_0)(z - z_0) \in F(z)\}.$$

By Corollary 2, Ψ is pseudo-Lipschitz around (y_0, z_0) if and only if the map $[F + f]^{-1}$ is pseudo-Lipschitz around $(y_0 + f(z_0), z_0)$.

In the remainder of the paper, we also make use of the following result.

COROLLARY 3. *If the assumptions of Corollary 2 hold, then for each $\gamma^+ > \gamma$, there exist positive constants a, ρ , and ε with the following properties: For every $z \in B_a(z_0)$, $p \in B_\rho(0)$, and $w_p \in F(z)$ with $\|T_p(z) - w_p\| \leq \varepsilon$, there exists $z_p \in \Sigma(p)$ such that*

$$(16) \quad \|z - z_p\| \leq \gamma^+ \|T_p(z) - w_p\|.$$

Proof. Choose σ and ρ as in the proof of Corollary 2 to satisfy (14). Let a and $\varepsilon > 0$ be small enough that

$$(17) \quad \frac{\rho}{2} \geq a \geq \gamma^+ \varepsilon \quad \text{and} \quad \sigma \geq \rho \bar{D}_\rho + \varepsilon.$$

Given $p \in B_\rho(0)$, $z \in B_a(z_0)$, and $w_p \in F(z)$ with $\|T_p(z) - w_p\| \leq \varepsilon$, let us define $y_p = w_p - L(z)$. We apply Theorem 1 with y_0 , z_0 , and T replaced by y_p , z , and T_p , and with $\delta = \|T_p(z) - w_p\|$. If $\delta = 0$, then $z \in \Sigma(p)$, and (16) holds. If $\delta > 0$, we take $r = \gamma^+ \delta$. Since $\delta \leq \varepsilon$, it follows from (14) and (17) that

$$\frac{\gamma \delta}{1 - \gamma E_\rho(p)} < r = \gamma^+ \delta \leq \gamma^+ \varepsilon \leq a \leq \frac{\rho}{2} < \rho.$$

Hence, $B_r(z) \subset B_\rho(z_0)$, which implies that $\Delta_r(p, z) \subset \Delta_\rho(p, z_0)$. By Remark 1 and (17), we have $\Delta_\rho(p, z_0) \subset B_\sigma(y_0)$ so that assumption (6) of Theorem 1 holds. By Theorem 1, there exists a solution $z_p \in B_r(z)$ to (11), where $r = \gamma^+ \delta$, which establishes (16). \square

Remark 3. Relation (16) implies that

$$\|z - \Sigma(p)\| \leq \gamma^+ \|T_p(z) - w_p\|.$$

Since the left-hand side of this inequality does not depend on w_p , it follows from Corollary 3 that when $\|T_p(z) - F(z)\|$ is sufficiently small, we have

$$(18) \quad \|z - \Sigma(p)\| \leq \gamma^+ \|T_p(z) - F(z)\|.$$

In particular, if $T_p(z) = T(z) + p$, we conclude that the map $T - F$ is *metrically regular* around $(z_0, 0)$. It turns out that metric regularity is equivalent to the pseudo-Lipschitz property (see Penot [33]). For a discussion of related results, see Cominetti [10], and the references therein.

Corollary 3 is a generalization of Theorem 1 in [36] where the estimate (18) is obtained under the following conditions: F is a closed, convex cone, independent of z ; $T_p(z)$ is continuously Fréchet differentiable; and interior point regularity holds. This regularity condition implies, via the celebrated Robinson-Ursescu theorem (see [36] and [43]), that the solution map of the linearized (auxiliary) problem is pseudo-Lipschitz.

3. Feasibility and controllability. As a first application of the abstract theory, we study the continuity of the map “parameter \rightarrow feasible set” of a nonlinear control system with constraints. The following model problem is analyzed: Given an interval $I = [0, 1]$, the state x is a map from I to R^n , while the control u is a map from I to R^m . Given θ between 1 and ∞ , let $L^\theta(R^m)$ denote the space of functions $u: I \rightarrow R^m$ with $|u(t)|^\theta$ integrable where $|\cdot|$ is the Euclidean norm. Let $W^{1,\theta}(R^n)$ denote the space of functions $x: I \rightarrow R^n$ with both x and its derivative in $L^\theta(R^n)$. We often omit the argument R^n or R^m when the context is clear. Of course, when θ is ∞ , these spaces are modified in the standard fashion: L^∞ is the space of essentially bounded functions, and $W^{1,\infty}$ is the space of Lipschitz continuous functions (or, equivalently, the space of essentially bounded functions with essentially bounded derivatives). Given functions

$$f_p: R^{n+m} \times I \rightarrow R^n, \quad K_p: R^m \times I \rightarrow R^\mu, \quad \text{and} \quad S_p: R^n \times I \rightarrow R^\nu,$$

where p is a parameter, and, given a starting condition $a \in R^n$, the feasible set $\Sigma(p)$ consists of the set of $u \in L^\infty$ and $x \in W^{1,\theta}$ that satisfy the relations

$$(19) \quad \begin{aligned} \dot{x}(t) &= f_p(x(t), u(t), t) \quad \text{and} \quad K_p(u(t), t) \leq 0 \quad \text{a.e. } t \in I, \\ x(0) &= a, \quad S_p(x(t), t) \leq 0 \quad \text{for every } t \in I. \end{aligned}$$

Using the notation of (11), the feasible set in the control problem consists of those $z \in \Omega$ such that $T_p(z) \in F(z)$, where

$$\Omega = \{z = (x, u): x \in W^{1,\theta}, u \in L^\infty, x(0) = a\},$$

$$T_p(x, u) = \begin{bmatrix} f_p(x, u) - \dot{x} \\ K_p(u) \\ S_p(x) \end{bmatrix}, \quad \text{and} \quad F(x, u) = \begin{bmatrix} 0 \\ L_-^\infty \\ L_-^\infty \end{bmatrix}.$$

Here L_-^∞ denotes the nonpositive functions in L^∞ .

Given a pair $z_0 = (x_0, u_0)$ that is feasible for (19) when $p = 0$, we wish to study the behavior of $\Sigma(p)$ for p near zero. Throughout this section, we make the following assumption: There exists a closed set $\Delta \subset R^n \times R^m \times I$ and a $\delta > 0$ such that $(x_0(t), u_0(t), t)$ lies in Δ for almost every $t \in I$, the distance from $(x_0(t), u_0(t), t)$ to the boundary of Δ in the hyperplane $R^n \times R^m \times \{t\}$ is at least δ for almost every $t \in I$, the derivatives of $f_p(x, u, t)$, $K_p(u, t)$, and $S_p(x, t)$ with respect to x and u exist on Δ , and these derivatives along with the function values are continuous with respect to $(x, u, t) \in \Delta$ and p near zero. From the development in § 2, Lipschitz properties of the solution map for the nonlinear problem are related to Lipschitz properties of the solution map for an auxiliary problem (4) when y is in a neighborhood of $T_0(z_0)$. We consider the following linearization of (19):

$$\begin{aligned} \dot{x}(t) - \dot{x}_0(t) &= A(t)(x(t) - x_0(t)) + B(t)(u(t) - u_0(t)) + y_1(t), \\ (20) \quad K(t)(u(t) - u_0(t)) + y_2(t) &\leq 0, \\ S(t)(x(t) - x_0(t)) + y_3(t) &\leq 0, \end{aligned}$$

where $y_1 \in L^\theta$, y_2 and $y_3 \in L^\infty$, and

$$A(t) = \nabla_x f_0(x_0(t), u_0(t), t),$$

$$B(t) = \nabla_u f_0(x_0(t), u_0(t), t),$$

$$K(t) = \nabla_u K_0(u_0(t), t),$$

$$S(t) = \nabla_x S_0(x_0(t), t).$$

Above any equality or inequality involving measurable functions is interpreted in the sense "almost everywhere."

From the development of § 2, we see that pseudo-Lipschitz continuity of the feasible map can be deduced from the following three conditions:

- (i) Lipschitz continuity of $T_p(z)$ with respect to p ,
- (ii) \bar{D}_ρ is sufficiently small,

(iii) The solution map associated with the linearized system is pseudo-Lipschitz. With regard to condition (i), Lipschitz continuity of $T_p(z)$ with respect to p is equivalent to Lipschitz continuity of $f_p(z)$, $K_p(z)$, and $S_p(z)$ with respect to p . Also, \bar{D}_ρ tends to zero as ρ tends to zero under our smoothness assumptions. In the following two lemmas, we study the Lipschitz continuity of the solution map for the linearized system.

LEMMA 2. *Let $z_0 = (x_0, u_0) \in \Omega$ be feasible in (19) when $p = 0$, let $\Lambda(y)$ denote the set of solutions $(x, u) \in W^{1,\theta} \times L^\infty$ to (20), and define $y_0 = T_0(z_0)$. If there exist $\alpha > 0$,*

$w \in W^{1,\theta}$, and $v \in L^\infty$ such that

$$(21) \quad \begin{aligned} \dot{w}(t) &= A(t)w(t) + B(t)v(t), & w(0) &= 0, \\ (K(t)v(t) + K_0(u_0(t), t))_i &\leq -\alpha, & i &= 1, 2, \dots, \mu, \\ (S(t)w(t) + S_0(x_0(t), t))_i &\leq -\alpha, & i &= 1, 2, \dots, \nu, \end{aligned}$$

then Λ is pseudo-Lipschitz around (y_0, z_0) .

Proof. This result follows from the Robinson-Ursescu theorem (see [36] and [43]) as stated (for example) by Aubin and Ekeland in [5, p. 132] or Clarke [9, p. 236]. That is, if there exists $r > 0$ such that for each y in a neighborhood of y_0 , we can find $z \in B_r(z_0)$ with $z \in \Lambda(y)$, then the map $y \rightarrow \Lambda(y) \cap B_r(z_0)$ is Lipschitz continuous, from which it follows that Λ is pseudo-Lipschitz around (y_0, z_0) . The proof of the lemma proceeds as follows: Given $y \in Y$ and a pair (w, v) satisfying (21), let x denote the solution to the differential equation in (20) corresponding to the control $u = v + u_0$ and the starting condition $x(0) = a$. Observe that x can be expressed as

$$x = w + x_0 + My_1,$$

where M is a bounded linear map from L^θ to $W^{1,\theta}$. Hence, we have

$$S(x - x_0) + y_3 = S(w + My_1) + y_3 \leq -\alpha + SM y_1 + y_3 - S_0(x_0).$$

Similarly, putting $u = v + u_0$ into the control constraint of (20) gives

$$K(u - u_0) + y_2 \leq -\alpha + y_2 - K_0(u_0).$$

Thus there exists $\sigma > 0$ such that x and u satisfy the constraints in (20) whenever $y \in B_\sigma(y_0)$, where

$$y_0 = T_0(z_0) = \begin{bmatrix} 0 \\ K_0(u_0) \\ S_0(x_0) \end{bmatrix}.$$

By the triangle inequality, we have

$$\|u - u_0\| + \|x - x_0\| \leq \|v\| + \|w\| + \|My_1\| \leq \|v\| + \|w\| + \sigma \|M\|$$

for every $y \in B_\sigma(y_0)$. Setting $r = \|v\| + \|w\| + \sigma \|M\|$, it follows that $B_r(z_0) \cap \Lambda(y)$ is nonempty whenever $y \in B_\sigma(y_0)$. This completes the proof. \square

The proof of Lemma 2 provides a way to construct a single-valued Ψ : For each $y \in Y$, x is the solution of the differential equation (20) corresponding to $u = v + u_0$. Now we present a condition that yields the existence of a state and control satisfying the interior point condition of Lemma 2. This condition is the same one that appeared in the study [18] of Lipschitz continuous solutions in optimal control. First, we provide some terminology. We say that a function g is piecewise continuous if there exists a finite sequence $\{t_i\}$ with

$$0 = t_0 < t_1 < t_2 < \dots < t_N = 1,$$

such that g is continuous on the open interval (t_i, t_{i+1}) for each i , and one-sided limits exist at each t_i . A function is piecewise continuously differentiable if it is continuous and its derivative is piecewise continuous. If $K(t)$ is the coefficient matrix for u in (20), then $K^B(t)$ and $K^N(t)$ denote the submatrices of $K(t)$ consisting of rows associated with those indices i for which either

$$K_0(u_0(t), t)_i = 0 \quad \text{or} \quad K_0(u_0(t), t)_i < 0, \quad \text{respectively.}$$

In other words, $K^B(t)$ and $K^N(t)$ are the submatrices corresponding to the binding and the nonbinding constraints at time t . The submatrices S^B and S^N of S are defined in a similar fashion. Basically, we will show that if the columns of $K^B(t)^T$ and

$B(t)^T S^B(t)^T$ are uniformly linearly independent, then the interior point condition of Lemma 2 is satisfied.

LEMMA 3. Let $z_0 = (x_0, u_0) \in \Omega$ be a point feasible for (19) when $p = 0$, and suppose that $S_0(a, 0) < 0$, $S_0(x_0)$ is piecewise continuously differentiable, both $K_0(u_0)$ and the matrices K and B are piecewise continuous, and at each t where these piecewise continuous functions are continuous, the following independence condition holds: There exists $\beta > 0$ such that

$$|K^B(t)^T b + B(t)^T S^B(t)^T c| \geq \beta(|b| + |c|)$$

for every b and c . Moreover, at a time t of discontinuity, this independence condition also holds, but with t replaced by both t^+ and t^- , and with the binding constraint set replaced by those of $K_0(u_0(t^+))$ and $K_0(u_0(t^-))$, respectively. Then there exist $w \in W^{1,\infty}$ and $v \in L^\infty$ that satisfy hypothesis (21) of Lemma 2 for some $\alpha > 0$.

Proof. Before considering the state constraints, we give a proof in the case of no state constraints. We first show that there exist a scalar $\delta > 0$ and sequences $\{t_i\}$ and $\{\tau_i\}$ such that

$$(22) \quad t_i \leq \tau_i \leq t_{i+1} \quad \text{for } 0 \leq i \leq N, \quad \tau_0 = t_0 = 0, \quad \tau_N = t_{N+1} = 1,$$

$$K_0^{N_i}(u_0(t), t) \leq -\delta \quad \text{for each } t_i \leq t \leq t_{i+1},$$

$$(23) \quad |K^{B_i}(t)^T b| \geq \delta |b| \quad \text{for each } t_i \leq t \leq t_{i+1}, \quad \text{for every } b,$$

where the B_i superscript means those rows (or components) associated with the binding constraints at τ_i , while the N_i superscript means those rows (or components) associated with nonbinding constraints at τ_i . The right-hand side of the inequality $K_0^{N_i}(u_0(t), t) \leq -\delta$ is interpreted as a vector with every component equal to $-\delta$.

To prove (22) and (23), we verify that they are satisfied on the closure of each open interval J where K and $K_0(u_0)$ are continuous. Let us define the parameter $\varepsilon(t)$ by

$$\varepsilon(t) = \text{minimum}_{1 \leq i \leq \mu} \{-K_0(u_0(t), t)_i : K_0(u_0(t), t)_i \neq 0\}.$$

If all the constraints are binding at t , then we set $\varepsilon(t) = +\infty$. The value of $K_0(u_0)$ at an endpoint of J is taken to be its limit at that point. By the continuity assumptions, it follows that for each $t \in J$, there exists an open ball O_t , containing t , such that

$$K_0^{N_i}(u_0(s)) \leq -\frac{\varepsilon(t)}{2} \quad \text{for every } s \in O_t$$

and

$$|K^{B_i}(s)^T b| \geq \frac{\beta}{2} |b| \quad \text{for every } s \in O_t,$$

where the superscript B_i stands for rows binding at t , while the superscript N_i stands for components nonbinding at t . If t is an endpoint of J , then the open ball is replaced by a half-open ball. By compactness, this cover of J has a finite subcover. In (22) and (23), the τ_i are the centers of the balls in the subcover, while the t_i are arbitrary points in the overlap region between adjacent balls. The parameter δ is given by

$$\delta = \frac{1}{2} \text{minimum} \{\varepsilon(\tau_0), \varepsilon(\tau_1), \dots, \varepsilon(\tau_N), \beta\}.$$

The control v that satisfies the interior point condition of Lemma 2 can be constructed in the following way: Given $\sigma > 0$ and t between t_i and t_{i+1} , $v(t)$ is the minimum norm solution of the equation $K^{B_i}(t)v(t) = -\sigma$ (if there are no binding constraints, set $v(t) = 0$). Since $K_0(u_0(t), t)$ is nonpositive, we have

$$(24) \quad K^{B_i}(t)v(t) + K_0^{B_i}(u_0(t), t) \leq -\sigma.$$

By (23), the smallest singular value of $K^{B_i}(t)$ is bounded from below by δ . It follows that $v(t)$ has the following bound in the Euclidean norm:

$$|v(t)| \leq \frac{\sigma \sqrt{\mu}}{\delta},$$

where μ is the number of component of K_0 . Hence, by (22) v also satisfies the inequality

$$(25) \quad K^{N_i}(t)v(t) + K_0^{N_i}(u_0(t), t) \leq \frac{\sigma \sqrt{\mu}}{\delta} |K^{N_i}(t)| - \delta.$$

Relations (24) and (25) imply that v satisfies the interior point condition (21) for σ sufficiently small.

When state constraints are present, this proof must be modified in several ways. By incorporating the state constraints in the definition of $\varepsilon(t)$, we can choose δ to satisfy the additional relation

$$S_0^{N_i}(x_0(t), t) \leq -\delta \quad \text{for each } t_i \leq t \leq t_{i+1}.$$

In addition, the independence condition (23) generalizes to the form

$$(26) \quad |K^{B_i}(t)^T b + B(t)^T S^{B_i}(t)^T c| \geq \delta(|b| + |c|).$$

Similar to the control constrained case, we wish to construct a control v_σ and a state w_σ such that $K^{B_i}v_\sigma(t) = -\sigma$ and $S^{B_i}w_\sigma(t) = -\sigma$ for t between t_i and t_{i+1} , and both v_σ and w_σ are bounded pointwise by a constant times σ . This construction is complicated by the fact that v_σ and w_σ must satisfy the linear differential equation in (21).

The proof proceeds by induction, interval by interval, from left to right. Suppose that on the interval $[t_0, t_k]$ we can construct a control v_σ and a corresponding state w_σ such that for each σ sufficiently small, we have

$$|w_\sigma(t_k)| \leq c\sigma, \quad S^{B_{k-1}}w_\sigma(t_k) = -\sigma,$$

where c is independent of σ , and with $\alpha = \sigma$, the control $v = v_\sigma$ and the state $w = w_\sigma$ satisfy the relations (21) on the interval $[0, t_k]$. We now show that this construction can be continued on the interval $[t_k, t_{k+1}]$. The control v and the state w on the new interval are chosen to satisfy the relations

$$(27) \quad \begin{aligned} K^{B_k}(t)v(t) &= -\sigma, \quad \text{for } t_k < t \leq t_{k+1}, \\ (S^{B_k}(t)w(t))_i &= (S^{B_k}(t_k)w_\sigma(t_k))_i + \gamma_i(t - t_k) \quad \text{for } t_k < t \leq t_k + \varepsilon, \\ S^{B_k}(t)w(t) &= -\sigma \quad \text{for } t_k + \varepsilon < t \leq t_{k+1}. \end{aligned}$$

The parameters γ_i and ε are selected so that $(S^{B_k}(t_k + \varepsilon)w(t_k + \varepsilon))_i = -\sigma$, or equivalently, so that

$$\gamma_i = -\frac{(S^{B_k}(t_k)w_\sigma(t_k))_i + \sigma}{\varepsilon}.$$

Since $w_\sigma(t_k)$ tends to zero as σ tends to zero, it follows that for any ε , γ_i tends to zero as σ tends to zero.

To obtain a control that satisfies (27), we differentiate the second and third equations in (27) and we substitute from the state equation $\dot{w} = Aw + Bv$ to obtain

$$(28) \quad \begin{aligned} S^{B_k}(t)B(t)v(t) &= \gamma_i - \frac{dS^{B_k}(t)}{dt} w(t) - S^{B_k}(t)A(t)w(t) \quad \text{for } t_k < t \leq t_k + \varepsilon, \\ S^{B_k}(t)B(t)v(t) &= -\frac{dS^{B_k}(t)}{dt} w(t) - S^{B_k}(t)A(t)w(t) \quad \text{for } t_k + \varepsilon < t \leq t_{k+1}. \end{aligned}$$

By the independence condition (26), the minimum norm control $v(t)$ that satisfies the equation $K^{B_k}(t)v(t) = -\sigma$ along with (28), where w is the solution to

$$\dot{w}(t) = A(t)w(t) + B(t)v(t), \quad w(t_k) = w_\sigma(t_k),$$

is bounded pointwise by a constant times σ . If the i th state constraint is binding at both τ_{k-1} and τ_k , then by the construction of w , we have

$$(S(t)w(t))_i = -\sigma \quad \text{for } t_k \leq t \leq t_{k+1},$$

which implies that

$$(S(t)w(t) + S_0(x_0(t), t))_i \leq -\sigma \quad \text{for } t_k \leq t \leq t_{k+1}.$$

If the i th state constraint is nonbinding at either τ_{k-1} or τ_k , then $S_0(x_0(t_k), t_k)_i \leq -\delta$. Hence, for ε sufficiently small, $S_0(x_0(t), t)_i \leq -\delta/2$ for t between t_k and $t_k + \varepsilon$. Taking σ sufficiently small yields

$$(S(t)w(t) + S_0(x_0(t), t))_i \leq -\delta/4 \quad \text{for } t_k \leq t \leq t_k + \varepsilon.$$

Now consider t in the interval $[t_k + \varepsilon, t_{k+1}]$. If the i th constraint is binding at τ_k , then

$$(S(t)w(t) + S_0(x_0(t), t))_i \leq -\sigma.$$

If the i th constraint is nonbinding at τ_k , then

$$(S(t)w(t) + S_0(x_0(t), t))_i \leq |S(t)w(t)| - \delta.$$

Since w is bounded by a constant times σ , the induction step is complete. \square

To conclude, we state a specific sensitivity result for the feasibility problem (19) based on Corollary 3.

THEOREM 2. *If (x_0, u_0) is feasible in (19) when $p = 0$, and there exist $\alpha > 0$, $w \in W^{1,\theta}$, and $v \in L^\infty$ satisfying (21), then for each p in a neighborhood of 0 and for each (x, u) in a neighborhood of $z_0 = (x_0, u_0)$, there exists x_p and u_p that are feasible in (19), and we have*

$$(29) \quad \|x_p - x\|_{W^{1,\theta}} + \|u_p - u\|_{L^\infty} \leq c(\|f_p(x, u) - \dot{x}\|_{L^\theta} + \|K_p(u)_+\|_{L^\infty} + \|S_p(x)_+\|_{L^\infty}),$$

where the “+” subscript stands for the positive part and c is independent of p .

Proof. Relation (29) follows from Corollary 3 and Lemma 2, where we identify the z of Corollary 3 with the pair (x, u) , while w_p is identified with the triple $(0, K_p(u)_-, S_p(x)_-)$. Here the subscript “-” stands for the negative part. \square

Remark 4. Using generalized derivatives of set-valued maps, a result related to Theorem 2 is established in [16, Thm. 10.1] for a problem with final state constraints. A linear system with convex state and control constraints is studied in [14].

We can use the same approach to study local controllability. The following simple example illustrates the basic ideas. Let us consider the nonsmooth control system

$$(30) \quad \dot{x}(t) = |x(t)| + x(t)^5 u(t) + u(t) \quad \text{a.e. } t \in I,$$

with the constraints

$$x(0) = 0, \quad x(t) \geq t - 1 \quad \text{for every } t \in I, \quad -1 \leq u(t) \leq 1 \quad \text{a.e. } t \in I, \quad x \in W^{1,\infty}, \quad u \in L^\infty.$$

A control system is *locally controllable* around 0 at $t = 1$ if for each a near zero that satisfies the state constraints at $t = 1$, there exists a feasible trajectory with $x(1) = a$. We will apply Corollary 1 with the following identifications:

$$\Omega = \{(x, u) \in W^{1,\infty} \times L^\infty : x(0) = 0, x(t) \geq t - 1 \text{ for every } t \in I, -1 \leq u(t) \leq 1 \text{ a.e. } t \in I\},$$

$$T_p(x, u) = \begin{bmatrix} \dot{x} - |x| - x^5 u - u \\ x(1) - p \end{bmatrix}, \quad \text{and} \quad F_p(x, u) = 0.$$

Local controllability is equivalent to existence of a solution to (7) for every $p \geq 0$, p sufficiently small.

In applying Corollary 1, we take $P = R_+$, the nonnegative real numbers, $z_p = z_0 = (x_0, u_0) = 0$, $y_0 = 0$, and

$$L_p(z) = \begin{bmatrix} \dot{x} - |x| - u \\ x(1) \end{bmatrix}.$$

With this definition, the auxiliary problem becomes the following:

$$(31) \quad \text{Find } (x, u) \in \Omega \text{ such that} \\ \dot{x}(t) - |x(t)| - u(t) + y_1(t) = 0, \quad x(1) + y_2 = 0.$$

Hypothesis (9) of Corollary 1 is satisfied if there exists a single-valued map Ψ from $L^\infty \times R$ to $W^{1,\infty} \times L^\infty$, with $\Psi(0) = 0$, with $(x, u) = \Psi(y_1, y_2)$ a solution of (31), and with Ψ Lipschitz continuous on the set $\Delta_\sigma(p)$ of Corollary 1. It can be verified that the following choice for Ψ has the desired properties:

$$\Psi(y) = \begin{bmatrix} \frac{1-e'}{e-1} y_2 \\ \frac{y_2}{1-e} - y_1(t) \end{bmatrix}.$$

(Note that if $(y_1, y_2) \in \Delta_\sigma(p)$, then $y_2 \leq 0$ since $p \geq 0$.) Hence, the control system (30) is locally controllable around 0 at $t = 1$.

4. Quadratic programs. In applying the results of § 2 to problems in optimal control and mathematical programming, we must derive Lipschitz results for the auxiliary problem. This section collects properties of quadratic programs that are relevant to the analysis.

LEMMA 4. *Let X denote a reflexive Banach space, let $\Lambda \subset X$ be a nonempty closed, convex subset, and consider the problem*

$$(32) \quad \text{minimize } \frac{1}{2} \langle Ax, x \rangle + \langle \phi, x \rangle \quad \text{over } x \in \Lambda,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between X and the dual space X^* , $\phi \in X^*$, $A: X \rightarrow X^*$ is a continuous linear operator, and $\langle Ax, y \rangle = \langle Ay, x \rangle$ for every x and $y \in X$. If there exists a constant $\alpha > 0$ such that

$$(33) \quad \langle A(x_1 - x_2), x_1 - x_2 \rangle \geq \alpha \|x_1 - x_2\|^2 \quad \text{for every } x_1 \text{ and } x_2 \in \Lambda,$$

then there is a unique solution \bar{x} to (32), and \bar{x} is the unique solution to the following variational inequality:

$$(34) \quad \text{Find } \bar{x} \in \Lambda \text{ such that } \langle A\bar{x} + \phi, x - \bar{x} \rangle \geq 0 \quad \text{for every } x \in \Lambda.$$

If x_1 and x_2 denote the solutions of (32) corresponding to $\phi = \phi_1$ and $\phi = \phi_2$, then we have

$$(35) \quad \|x_1 - x_2\| \leq \|\phi_1 - \phi_2\| / \alpha.$$

Proof. In a Hilbert space, the existence of a solution \bar{x} to (32) and the correspondence between a solution to (32) and a solution to (34) is found for example, in [26, Chap. 1]. The Lipschitz result (35) is found in [20, Lemma 1] for a Hilbert space. These Hilbert space proofs are also valid in a reflexive Banach space. \square

The usual second-order sufficient condition for (32) has the form

$$(36) \quad \langle A(x - \bar{x}), x - \bar{x} \rangle \geq \alpha \|x - \bar{x}\|^2 \quad \text{for every } x \in \Lambda,$$

where $\alpha > 0$. Hence, condition (33) is stronger than the second-order sufficient condition. An important difference between (33) and (36) is that after small perturbations in A , (33) still holds for some $\alpha > 0$; after small perturbations in A and \bar{x} , (36) may not hold for any $\alpha > 0$.

Under the hypotheses of Lemma 4, let us consider the constraint set

$$(37) \quad \Lambda = \{x \in X: Bx + \psi \in K\},$$

where $B: X \rightarrow W$ is a continuous, linear operator; W is a Banach space; $\psi \in W$; and $K \subset W$ is a closed, convex cone with vertex at the origin. In this case, (32) takes the form

$$(38) \quad \text{minimize } \frac{1}{2}\langle Ax, x \rangle + \langle \phi, x \rangle \quad \text{subject to } Bx + \psi \in K.$$

Given x_1 and $x_2 \in \Lambda$, observe that $v = x_1 - x_2$ has the property that $Bv \in K - K$. Hence, when Λ is given by (37), (33) holds if

$$(39) \quad \langle Av, v \rangle \geq \alpha \|v\|^2 \quad \text{whenever } Bv \in K - K.$$

Conversely, if B is surjective and (33) holds, then (39) holds. Hence, (39) and (33) are equivalent when B is surjective.

Letting K^+ denote the polar cone defined by

$$K^+ = \{\lambda \in W^*: \langle \lambda, k \rangle \geq 0 \text{ for every } k \in K\},$$

suppose that there exists $\lambda \in K^+$ and $\bar{x} \in X$ satisfying

$$(40) \quad A\bar{x} - B^*\lambda + \phi = 0 \quad \text{and} \quad \langle \lambda, B\bar{x} + \psi \rangle = 0, \quad \text{where } B\bar{x} + \psi \in K.$$

It follows that

$$0 = \langle A\bar{x} + \phi, x - \bar{x} \rangle - \langle \lambda, Bx + \psi \rangle \leq \langle A\bar{x} + \phi, x - \bar{x} \rangle$$

for every $x \in X$ with $Bx + \psi \in K$. By Lemma 4, \bar{x} is the unique solution to (32). Note that the conditions $\langle \lambda, B\bar{x} + \psi \rangle = 0$ and $B\bar{x} + \psi \in K$ of (40) are often written in the compact form

$$B\bar{x} + \psi \in \partial K^+(\lambda),$$

where $\partial K^+(\lambda) = \{w \in W^{**}: \langle w, \mu - \lambda \rangle \geq 0 \text{ for each } \mu \in K^+\}$ is the normal cone at λ to the set K^+ .

If \bar{x} is a solution to (38) and B is surjective, it is known (see Kurcyusz [25]) that there exists a unique Lagrange multiplier $\lambda \in K^+$ satisfying (40). Assuming that B is surjective and (33) holds, let us study the dependence of the solution and the multiplier associated with (38) on the parameters ϕ and ψ . Given $\phi = \phi_i \in X^*$ and $\psi = \psi_i \in W$ for $i = 1$ and 2 , let \bar{x}_i be the corresponding solutions to (38), and let λ_i be the associated multipliers satisfying (40). If $x = \tilde{x}_1$ is any solution to $Bx = -\psi_1$, then by the open mapping principle (see [7, p. 57]), B^{-1} is Lipschitz continuous, and there exists a solution $x = \tilde{x}_2$ to $Bx = -\psi_2$ such that

$$\|\tilde{x}_1 - \tilde{x}_2\| \leq c \|\psi_1 - \psi_2\|,$$

where c is independent of ψ_1 and ψ_2 . Making the change of variables $\bar{x}_i = w_i + \tilde{x}_i$ in (40), we obtain

$$(41) \quad Aw_i - B^*\lambda_i + A\tilde{x}_i + \phi_i = 0, \quad Bw_i \in K, \quad \text{and} \quad \langle \lambda_i, Bw_i \rangle = 0.$$

Hence, $w = w_i$ is the solution to the problem

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\langle Aw, w \rangle + \langle A\tilde{x}_i + \phi_i, w \rangle \\ &\text{subject to} \quad Bw \in K. \end{aligned}$$

By Lemma 4, inequality (35), we have

$$\alpha \|w_1 - w_2\| \leq \|\phi_1 - \phi_2\| + \|A\| \|\tilde{x}_1 - \tilde{x}_2\| \leq \|\phi_1 - \phi_2\| + c \|A\| \|\psi_1 - \psi_2\|.$$

Taking the norm of the identity $x_1 - x_2 = w_1 - w_2 + \tilde{x}_1 - \tilde{x}_2$, and applying the triangle inequality, we obtain

$$\|x_1 - x_2\| \leq c \|\phi_1 - \phi_2\| + c \|\psi_1 - \psi_2\|,$$

where c is independent of the ϕ_i and the ψ_i . Moreover, since B^* is one-to-one and $(B^*)^{-1}$ is a continuous linear operator, (40) implies that $\|\lambda_1 - \lambda_2\|$ has a similar bound. These observations are summarized in the following lemma.

LEMMA 5. *Suppose that B is surjective and (33) holds. If x_i and λ_i are the solutions to (40) corresponding to $\phi = \phi_i$ and $\psi = \psi_i$, $i = 1$ and 2 , then there exists a constant c , depending only on A and B , such that*

$$\|x_1 - x_2\| + \|\lambda_1 - \lambda_2\| \leq c \|\phi_1 - \phi_2\| + c \|\psi_1 - \psi_2\|.$$

Clearly, the coercivity condition (33) is preserved after small perturbations in A . In the context of (38) with B surjective, we now show that the coercivity condition (33) is preserved after small perturbations in B , and after arbitrary perturbations in ψ . Since B is surjective, we observed earlier that (33) is equivalent to (39). Since ψ does not appear in (39), coercivity is preserved after any perturbation in ψ . Now let us consider the effect of changes in B . Given a bounded linear operator $\bar{B}: X \rightarrow W$, the open mapping principle implies that there exists a constant c , depending only on B , with the following property: If $\bar{B}\bar{v} \in K - K$, we can find $v \in X$ with $Bv = \bar{B}\bar{v} \in K - K$ and

$$(42) \quad \|v - \bar{v}\| \leq c \|B - \bar{B}\| \|\bar{v}\|.$$

By the triangle inequality, we have

$$(43) \quad (1 - c \|B - \bar{B}\|) \|\bar{v}\| \leq \|v\| \leq (1 + c \|B - \bar{B}\|) \|\bar{v}\|.$$

Defining $\delta v = v - \bar{v}$ and applying (33) yields

$$(44) \quad \langle A\bar{v}, \bar{v} \rangle \geq \alpha \|v\|^2 - 2\langle Av, \delta v \rangle + \langle A\delta v, \delta v \rangle.$$

Utilizing the inequality $2ab \leq \rho a^2 + b^2/\rho$, where ρ is an arbitrary scalar, we have

$$|\langle Av, \delta v \rangle| \leq \|A\| \|v\| \|\delta v\| \leq \frac{\alpha}{4} \|v\|^2 + \frac{\|A\|^2}{\alpha} \|\delta v\|^2.$$

This inequality, coupled with relations (42)–(44), yields the following result.

LEMMA 6. *If the coercivity condition (39) holds and B is surjective, then for \bar{A} in a neighborhood of A and for \bar{B} in a neighborhood of B , there exists $\alpha > 0$ such that*

$$\langle \bar{A}v, v \rangle \geq \alpha \|v\|^2 \text{ whenever } \bar{B}v \in K - K.$$

We observe that in certain cases, the coercivity condition (33) holds for the set Λ if it holds on a subset $\Lambda \cap \Gamma$.

LEMMA 7. *Let Γ and Λ be convex subsets of X , and suppose that there exists $\alpha > 0$ such that*

$$(45) \quad \langle A(x_1 - x_2), x_1 - x_2 \rangle \geq \alpha \|x_1 - x_2\|^2 \text{ for every } x_1 \text{ and } x_2 \in \Lambda \cap \Gamma.$$

If $\text{int } \Gamma$, the interior of Γ , intersects Λ , then (33) holds.

Proof. Given $v_i \in \Lambda$, $i = 1$ and 2 , and $v \in \text{int } \Gamma \cap \Lambda$, define $v_i(\beta)$ by

$$v_i(\beta) = \beta v_i + (1 - \beta)v.$$

Since Λ is convex, $v_i(\beta) \in \Lambda$ for $i = 1$ and 2 , whenever $\beta \in [0, 1]$. Since $v_i(\beta) \rightarrow v$ as $\beta \rightarrow 0$, we can choose $\beta > 0$ small enough that $v_i(\beta) \in \Gamma$ for $i = 1$ and 2 . Applying (45) with $x_i = v_i(\beta)$, we obtain (33). \square

Remark 5. Suppose that $B = (B_1, B_2)$, $\psi = (\psi_1, \psi_2)$, and $K = K_1 \times K_2$, where B_1 is surjective, and where there exists $v \in X$ with $B_1 v + \psi_1 \in K_1$ and $B_2 v + \psi_2 \in \text{int } K_2$. Combining Lemmas 6 and 7, we see that if the coercivity condition (33) holds, then for \bar{A} and \bar{B}_1 near A and B_1 respectively, there exists $\alpha > 0$ such that

$$\langle \bar{A}v, v \rangle \geq \alpha \|v\|^2 \quad \text{whenever } \bar{B}_1 v \in K_1 - K_1.$$

5. Optimal solutions. In this section, we use Corollaries 1 and 2 to study an optimal solution of the problem

$$(46) \quad \text{minimize } C_p(x) \quad \text{over } x \in \Omega_p,$$

where p is a parameter in a metric space P , Ω_p is a closed, convex nonempty subset of a reflexive Banach space X , and $C_p: X \rightarrow \mathbb{R}$. Given a local minimizer x_0 of (46) corresponding to $p = 0$, we assume that for each $p \in P$, the functional $C_p(x)$ is twice Fréchet differentiable with respect to x in a neighborhood of x_0 , the derivatives $C'_p(x)$ and $C''_p(x)$ with respect to x are continuous in p and x in a neighborhood of $p = 0$ and $x = x_0$, and there exists $\alpha > 0$ such that

$$(47) \quad \langle C''_0(x_0)(x_1 - x_2), x_1 - x_2 \rangle \geq \alpha \|x_1 - x_2\|^2 \quad \text{for every } x_1, x_2 \in \bigcup_{p \in P} \Omega_p.$$

In addition, we assume that $\lim_{p \rightarrow 0} |\Omega_p - \Omega_0| = 0$, where $|A - B| = \|A - B\| + \|B - A\|$ is equivalent to the Hausdorff distance between the sets A and B .

THEOREM 3. For each $\beta > 1/\alpha$ and $\gamma > \|C'_0(x_0)\|$, there exists $s > 0$ with the following property: For each $p \in B_s(0)$, we can find a strict local minimizer x_p of (46) such that

$$(48) \quad \|x_p - x_0\| \leq \beta \|C'_p(x_0) - C'_0(x_0)\| + \sqrt{\gamma/\alpha} |\Omega_p - \Omega_0|^{1/2}.$$

If $C'_0(x_0) = 0$, then we require $\gamma > \|C''_0(x_0)\|$, and we replace the exponent $\frac{1}{2}$ in (48) by 1.

Proof. Since x_0 is a local minimizer of (46) and Ω_0 is convex, we have

$$(49) \quad \langle C'_0(x_0), x - x_0 \rangle \geq 0 \quad \text{for every } x \in \Omega_0.$$

Given $p \in P$, (47) and Lemma 4 imply that there exists a unique $\xi_p \in \Omega_p$ satisfying the relation

$$(50) \quad \langle C'_0(x_0) + L(\xi_p - x_0), x - \xi_p \rangle \geq 0 \quad \text{for every } x \in \Omega_p,$$

where $L = C''_0(x_0)$. Adding (49) with $x = z_0$ and (50) with $x = z_p$ to inequality (47) with $x_1 = \xi_p$ and $x_2 = x_0$, we get

$$\alpha \|\xi_p - x_0\|^2 \leq \langle L(\xi_p - x_0), z_p - x_0 \rangle + \langle C'_0(x_0), z_0 - \xi_p + z_p - x_0 \rangle$$

for every $z_0 \in \Omega_0$ and $z_p \in \Omega_p$. From this, it follows that

$$\alpha \|\xi_p - x_0\|^2 \leq \|L\| \|\xi_p - x_0\| \|x_0 - \Omega_p\| + \|C'_0(x_0)\| (\|x_0 - \Omega_p\| + \|\xi_p - \Omega_0\|).$$

Consequently, $\xi_p \rightarrow x_0$ as $p \rightarrow 0$, and we have

$$(51) \quad \alpha \|\xi_p - x_0\|^2 \leq (\|L\| \|\xi_p - x_0\| + \|C'_0(x_0)\|) |\Omega_p - \Omega_0|.$$

Now let us consider the following problem:

$$(52) \quad \text{Find } x_p \in \Omega_p \text{ such that } \langle C'_p(x_p), x - x_p \rangle \geq 0 \quad \text{for every } x \in \Omega_p.$$

We apply Corollary 1 to this problem, making the following identifications:

$$z = x, \quad T_p = C'_p, \quad y_p = C'_0(x_0) - L(x_0), \quad \text{and} \quad F_p(x) = \partial\Omega_p(x).$$

The auxiliary problem is the following:

$$(53) \quad \text{Find } x \in \Omega_p \text{ such that } \langle L(x) + y, w - x \rangle \geq 0 \quad \text{for every } w \in \Omega_p.$$

By Lemma 4, there exists a unique solution of (53) for each $y \in X^*$, and this solution is a Lipschitz continuous function of y with Lipschitz constant $1/\alpha$, independent of p . By Corollary 1 and for any constant $\beta > 1/\alpha$, there exists a solution x_p to (52) for p near zero, and we have

$$(54) \quad \|x_p - \xi_p\| \leq \beta \|C'_p(\xi_p) - L(\xi_p) - y_p\|.$$

By the definition of y_p , it follows that

$$\|C'_p(\xi_p) - L(\xi_p) - y_p\| = \|C'_p(\xi_p) - C'_0(x_0) - L(\xi_p - x_0)\|.$$

By the continuous differentiability assumptions, it follows that for any $\varepsilon > 0$, there exists an $s > 0$ such that

$$\|C'_0(\xi_p) - C'_0(x_0) - L(\xi_p - x_0)\| \leq \varepsilon \|\xi_p - x_0\| \quad \text{for every } p \in B_s(0)$$

and

$$\|C'_p(\xi_p) - C'_0(\xi_p)\| \leq \|C'_p(x_0) - C'_0(x_0)\| + \varepsilon \|\xi_p - x_0\| \quad \text{for every } p \in B_s(0).$$

These inequalities, combined with (51), (54), and the triangle inequality, yield (48).

By (47), x_p is a strict local minimizer of (46) for p near 0. \square

Remark 6. In general, the exponent $\frac{1}{2}$ in (48) is sharp (see [13, p. 13]). Theorem 3 is a generalization of Proposition 1.2 in [13].

Typically, Theorem 3 yields a Hölder-type estimate for $x_p - x_0$. However, when the constraint set is described by equalities and inequalities that possess certain regularity properties, a Lipschitz estimate can be established. We consider the following problem:

$$(55) \quad \text{minimize } C_p(x) \quad \text{subject to } G_p(x) \in K,$$

where $G_p: X \rightarrow W$, W is a Banach space, and $K \subset W$ is a closed, convex cone with vertex at the origin. Letting x_0 denote a local minimizer of (55) corresponding to $p = 0$, we assume henceforth in this section that C_p and G_p possess the following smoothness properties: $C_p(x)$ and $G_p(x)$ are twice Fréchet differentiable in x in a neighborhood of $p = 0$ and $x = x_0$, and these derivatives are continuous in p and x at $p = 0$ and $x = x_0$. The functions $G_p(x)$, $C'_p(x)$, and $G'_p(x)$ are Lipschitz in $p \in P$, uniformly in x near x_0 . Let H_p denote the Lagrangian defined by

$$H_p(x, \lambda) = C_p(x) - \langle \lambda, G_p(x) \rangle,$$

where $\lambda \in W^*$. The first-order necessary conditions associated with a solution to (55) can be expressed in the following way:

$$(56) \quad \nabla_x H_p(x_p, \lambda_p) = 0 \quad \text{and} \quad G_p(x_p) \in \partial K^+(\lambda_p), \quad \text{where } x_p \in X \quad \text{and} \quad \lambda_p \in K^+.$$

It is well known (see [25]) that if $G'_0(x_0)$ is surjective, then there exists λ_0 satisfying (56) for $p = 0$. Our Lipschitz result is based on the following coercivity condition: There exists $\alpha > 0$ such that

$$(57) \quad \begin{aligned} & \langle \nabla_{xx}^2 H_0(x_0, \lambda_0)(x_2 - x_1), x_2 - x_1 \rangle \geq \alpha \|x_2 - x_1\|^2 \\ & \text{whenever } G_0(x_0) + G'_0(x_0)(x_i - x_0) \in K, \quad \text{for } i = 1, 2. \end{aligned}$$

THEOREM 4. *If $G'_0(x_0)$ is surjective and the coercivity condition (57) holds, then there exists $s > 0$ such that (55) has a strict local minimizer x_p for each $p \in B_s(0)$, and both x_p , and the associated (unique) multiplier $\lambda_p \in K^+$ satisfying the first-order necessary condition (56), are Lipschitz continuous functions of $p \in B_s(0)$.*

Proof. We apply Corollary 2 with the following identifications: $z = (x, \lambda)$, $Z = X \times X^*$, $\Omega = X \times K^+$,

$$T_p(x, \lambda) = \begin{bmatrix} \nabla_x H_p(x, \lambda) \\ G_p(x) \end{bmatrix}, \quad \text{and} \quad F(x, \lambda) = \begin{bmatrix} 0 \\ \partial K^+(\lambda) \end{bmatrix}.$$

Hence, the problem "Find $z \in \Omega$ such that $T_p(z) \in F(z)$ " is the same as finding x and λ satisfying the first-order necessary condition for (55). In the auxiliary problem, we take $L = T'_0(z_0)$. Hence, the auxiliary problem is equivalent to the following: Given $\phi \in X^*$ and $\psi \in W$, find $x \in X$ and $\lambda \in K^+$ such that

$$(58) \quad \begin{aligned} Ax - B^*\lambda + \phi &= 0 \quad \text{and} \quad Bx + \psi \in \partial K^+(\lambda), \\ \text{where } A &= \nabla_{xx}^2 H_0(x_0, \lambda_0) \quad \text{and} \quad B = G'_0(x_0). \end{aligned}$$

By Lemma 5, the solution to (58) is a Lipschitz continuous function of ϕ and ψ . By Corollary 2, there exists a solution $z_p = (x_p, \lambda_p)$ to (56), which is a Lipschitz continuous function of p for p near 0. By Lemma 6, the coercivity condition (57) holds when the zeros are replaced by p near 0. Hence, the second-order sufficiency condition holds (see Maurer and Zowe [30]), and x_p is a strict local minimizer of (55) for p near 0. \square

Theorem 4 yields Lipschitz continuity without assuming strict complementary slackness. For an illustration, suppose that $G_p = (g_p, h_p)$ and $K = K_g \times K_h$, where K_g and K_h are closed convex cones with vertices at the origin of the associated Banach spaces. In this cases, the optimization problem (55) takes the form

$$(59) \quad \begin{aligned} &\text{minimize} \quad C_p(x) \\ &\text{subject to} \quad g_p(x) \in K_g, \quad h_p(x) \in K_h. \end{aligned}$$

The Lagrangian H_p is given by

$$H_p(x, \mu, \nu) = C_p(x) - \langle \mu, g_p(x) \rangle - \langle \nu, h_p(x) \rangle,$$

where $\lambda = (\mu, \nu)$ is the multiplier in the dual space. Again, if x_0 is a local minimizer of (59) and $G'_0(x_0)$ is surjective, then there exists a multiplier $\lambda_0 = (\mu_0, \nu_0)$ satisfying (56) for $p = 0$. Let us assume that $h_0(x_0) = 0$ and $\nu_0 \in \text{int } K_h^+$. In finite dimensions, h_p corresponds to the part of the inequality constraints that are active at $p = 0$ with associated multipliers that are strictly positive. We make the following coercivity assumption:

$$(60) \quad \begin{aligned} &\langle \nabla_{xx}^2 H_0(x_0, \mu_0, \nu_0)(x_2 - x_1), x_2 - x_1 \rangle \geq \alpha \|x_2 - x_1\|^2 \\ &\text{whenever } g_0(x_0) + g'_0(x_0)(x_i - x_0) \in K_g, \quad h'_0(x_0)(x_i - x_0) = 0, \quad \text{for } i = 1, 2. \end{aligned}$$

By Theorem 4, the following optimization problem has a local minimizer for p near 0 that depends Lipschitz continuously on p :

$$(61) \quad \begin{aligned} &\text{minimize} \quad C_p(x) \\ &\text{subject to} \quad g_p(x) \in K_g, \quad h_p(x) = 0. \end{aligned}$$

Observe that this problem differs from (59) since the constraint $h_p(x) \in K_h$ associated with (59) has been replaced by $h_p(x) = 0$. Exploiting the assumption that v_0 lies in the interior of K_h^+ , we show that this local minimizer for (61) is also a local minimizer of (59).

COROLLARY 4. *If $(g'_0(x_0), h'_0(x_0))$ is surjective, the coercivity condition (60) holds, and v_0 lies in the interior of K_h^+ , then there exists $s > 0$ such that (59) has a strict local minimizer x_p for each $p \in B_s(0)$, and both x_p , and the associated multipliers $\mu_p \in K_g^+$ and $\nu_p \in K_h^+$ satisfying the first-order necessary condition, are Lipschitz continuous functions of $p \in B_s(0)$.*

Proof. We apply the proof given for Theorem 4 to problem (61) replacing K by $K_g \times \{0\}$ and replacing the coercivity condition (57) by (60). It follows that there exists a solution x_p of (61) and associated Lagrange multipliers μ_p and ν_p that are Lipschitz continuous functions of p near 0 and that satisfy the first-order necessary conditions for (61). Since $\nu_p \in \text{int } K_h^+$ for p near zero, the first-order necessary conditions for (59) hold. By Lemma 8 of Appendix 1, x_p is a strict local minimizer of (61). \square

Finally, let us suppose that $G_p = (f_p, g_p, h_p)$ and $K = K_f \times K_g \times K_h$, where K_f , K_g , and K_h are closed convex cones with vertices at the origin of the associated Banach spaces. Hence, the optimization problem (55) takes the form

$$(62) \quad \begin{array}{ll} \text{minimize} & C_p(x) \\ \text{subject to} & f_p(x) \in K_f, \quad g_p(x) \in K_g, \quad h_p(x) \in K_h. \end{array}$$

If $f_0(x_0) \in \text{int } K_f$, then under the hypotheses of Corollary 4, the solution x_p of (59) is a Lipschitz continuous function of p and $f_p(x_p) \in K_f$ for p near 0. Hence, the local minimizer x_p of (59) is a local minimizer of (62).

Remark 7. Theorem 4 and Corollary 4 yield Lipschitz continuity of a local minimizer in a neighborhood of a reference point, without assuming strict complementary slackness. In finite dimensions, this problem was studied by Hager in [18] and by Robinson in [37]. Note that the coercivity assumption (60) is slightly weaker, in the infinite-dimensional context, than the coercivity assumption used in earlier work (see [18], [23], [37]) since (60) only requires coercivity relative to those x_i satisfying the constraint $g_0(x_0) + g'_0(x_0)(x_i - x_0) \in K_g$.

In comparing Corollary 4 to the recent paper [23] of Ito and Kunisch, note that in [23] the infinite-dimensional constraints are linear inequalities, the problem is formulated in a Hilbert space, and the nonlinear constraints for which the associated dual multiplier can vanish are finite-dimensional. Alt presents in [1] and [3] a stability analysis that is related to ours, but different. In [1] he considers a cone constrained problem, under the assumption that any neighborhood of the reference point contains a solution of the perturbed problem. In [3] he studies Lipschitz continuity of the solution of a problem with nonlinear cone constraints and with equality constraints under a weaker constraint qualification (Robinson's constraint regularity condition), but a stronger coercivity—(60) is required to hold on the kernel of the gradients of the equality constraints; moreover, he assumes that the variation of the Lagrange multipliers for the perturbed problem can be estimated in terms of the variation in the solution and the variation in the parameter (under our surjectivity condition, this hypothesis is satisfied). Recently, Malanowski [28] has obtained a Lipschitz continuity result in a Hilbert space setting that parallels the analysis of Ito and Kunisch [23], using a regularity condition for the constraints that is weaker than surjectivity.

6. Optimal control. Let us consider a nonlinear optimal control problem with control constraints

$$(63) \quad \begin{aligned} & \text{minimize} \quad \int_I g_p(x(t), u(t)) dt \\ & \text{subject to} \quad \dot{x}(t) = f_p(x(t), u(t)) \quad \text{and} \quad u(t) \in U \quad \text{a.e. } t \in I, \\ & \quad \quad \quad x(0) = a, \quad x \in W^{1,\theta}, \quad u \in L^\infty, \end{aligned}$$

where $f_p: R^{n+m} \rightarrow R^n$, $g_p: R^{n+m} \rightarrow R$, $U \subset R^m$ is nonempty, closed, and convex, a is the given starting condition, and $\theta \in [1, \infty]$. We assume that there exists a solution (x_0, u_0) to (63) corresponding to $p = 0$, and we wish to show that there exists a nearby solution for p in a neighborhood of zero. To this end, suppose that there exists a closed set $\Delta \subset R^n \times R^m$ and a $\delta > 0$ such that $(x_0(t), u_0(t))$ lies in Δ for almost every $t \in I$, the distance from $(x_0(t), u_0(t))$ to the boundary of Δ is at least δ for almost every $t \in I$, and the first two derivatives of $f_p(x, u)$ and $g_p(x, u)$ with respect to x and u exist on Δ , and these derivatives, along with the function value $f_p(x, u)$, are continuous with respect to $(x, u) \in \Delta$ and p near zero.

Let H_p denote the Hamiltonian defined by

$$H_p(x, u, \lambda) = g_p(x, u) + \lambda^T f_p(x, u).$$

If (x_p, u_p) is a solution of (63), then the minimum principle [22, p. 134] implies the following:

$$\nabla_u H_p(x_p(t), u_p(t), \lambda_p(t))^T (v - u_p(t)) \geq 0 \quad \text{a.e. } t \in I \text{ and for every } v \in U,$$

where $\lambda = \lambda_p$ is the solution of the adjoint equation

$$\dot{\lambda}(t) = -\nabla_x H_p(x(t), u(t), \lambda(t)) \quad \text{a.e. } t \in I, \quad \lambda(1) = 0,$$

associated with $x = x_p$ and $u = u_p$. Let $f_0^*(t)$ and $H_0^*(t)$ stand for $f_0(x_0(t), u_0(t))$ and $H_0(x_0(t), u_0(t), \lambda_0(t))$, and define the matrices

$$\begin{aligned} A(t) &= \nabla_x f_0^*(t), \quad B(t) = \nabla_u f_0^*(t), \quad Q(t) = \nabla_{xx}^2 H_0^*(t), \\ R(t) &= \nabla_{uu}^2 H_0^*(t), \quad S(t) = \nabla_{xu}^2 H_0^*(t). \end{aligned}$$

The following coercivity assumption will be utilized: There exists $\alpha > 0$ such that

$$(64) \quad \int_I (x(t)^T Q(t)x(t) + u(t)^T R(t)u(t) + 2x(t)^T S(t)u(t)) dt \geq \alpha \int_I |u(t)|^2 dt$$

whenever $x \in W^{1,2}$, $x(0) = 0$, $u \in L^2$, $\dot{x} = Ax + Bu$, $u = v - w$ for some v and $w \in L^2$ with $v(t)$ and $w(t) \in U$ for almost every $t \in I$. By taking $v = w$ except on a small interval, it can be shown, below, that a pointwise coercivity condition holds (see the recent paper [15]):

$$(65) \quad u^T R(t)u \geq \alpha |u|^2 \quad \text{a.e. } t \in I \text{ whenever } u = v - w \text{ with } v \text{ and } w \in U.$$

THEOREM 5. *If the coercivity condition (64) holds, then there exists positive constants κ , r , and s such that for each $p \in B_s(0)$, (63) has a strict local minimizer $(x_p, u_p) \in B_r(x_0, u_0)$ and the relation*

$$\begin{aligned} & \|x_p - x_q\|_{W^{1,\theta}} + \|u_p - u_q\|_{L^\infty} + \|\lambda_p - \lambda_q\|_{W^{1,\theta}} \\ & \leq \kappa (\|f_q(x_p, u_p) - f_p(x_p, u_p)\|_{L^\theta} \\ & \quad + \|\nabla_u H_q(x_p, u_p, \lambda_p) - \nabla_u H_p(x_p, u_p, \lambda_p)\|_{L^\infty} \\ & \quad + \|\nabla_x H_q(x_p, u_p, \lambda_p) - \nabla_x H_p(x_p, u_p, \lambda_p)\|_{L^\theta}) \end{aligned}$$

holds whenever p and $q \in B_s(0)$.

Proof. We apply Corollary 2 where the z of Corollary 2 is identified with (x, u, λ) , while Ω , T_p , and F are defined in the following way:

$$\Omega = \{(x, u, \lambda): x \in W^{1,\theta}, u \in L^\infty, \lambda \in W^{1,\theta}, x(0) = a, \\ \lambda(1) = 0, u(t) \in U \text{ a.e. } t \in I\},$$

$$T_p(x, u, \lambda) = \begin{bmatrix} \nabla_x H_p(x, u, \lambda) + \dot{\lambda} \\ \nabla_u H_p(x, u, \lambda) \\ f_p(x, u) - \dot{x} \end{bmatrix}, \quad \text{and} \quad F(x, u, \lambda) = \begin{bmatrix} 0 \\ \partial U(u) \\ 0 \end{bmatrix}.$$

The space Y containing the range of T_p is $L^\theta \times L^\infty \times L^\theta$. We make the following choice for the operator L of Corollary 2: $L(z) = M(z - z_0)$, where

$$M(x, u, \lambda) = \begin{bmatrix} A^T \lambda + Qx + Su + \dot{\lambda} \\ Ru + S^T x + B^T \lambda \\ Ax + Bu - \dot{x} \end{bmatrix}.$$

It can be verified that under the smoothness assumptions, $E_r(p) \rightarrow 0$ as p and r tend to zero, where $E_r(p)$ is defined in (12).

Given q_i and s_i in L^θ and r_i in L^∞ for $i = 1$ and 2 , let us consider the following problem:

$$(66) \quad \text{Find } (x, u, \lambda) \in \Omega \text{ such that } L(x, u, \lambda) + \begin{bmatrix} q_i \\ r_i \\ s_i \end{bmatrix} \in F(x, u, \lambda).$$

In [20, Lemma 3], we show that when the coercivity assumptions (64) and (65) hold, (66) has a unique solution (x_i, u_i, λ_i) , and the following Lipschitz property holds:

$$\|x_2 - x_1\|_{W^{1,\theta}} + \|u_2 - u_1\|_{L^\infty} + \|\lambda_2 - \lambda_1\|_{W^{1,\theta}} \\ \leq \gamma[\|q_2 - q_1\|_{L^\theta} + \|r_2 - r_1\|_{L^\infty} + \|s_2 - s_1\|_{L^\theta}].$$

Hence, Ψ is Lipschitz, and by Corollary 2, problem (11) has a locally unique solution that satisfies the conclusion of Theorem 5. Since the estimate of Theorem 5 yields an L^∞ perturbation in both state and the control, and since the coercivity assumption (64) is preserved after small perturbations in Q , R , S , A , and B , it follows from Corollary 5 in Appendix 1 that the solution of (11) provided by Corollary 2 is a strict local minimizer for the optimal control problem (63) when p is near zero. \square

We show by an example that Lipschitz continuity can be obtained without the coercivity condition (64). Consider the following problem:

$$(67) \quad \begin{aligned} &\text{minimize} \quad x_1(1) + x_2(1) \\ &\text{subject to} \quad \dot{x}_1 = px_1 \sin x_2 + u_1, \quad \dot{x}_2 = u_2, \quad x_1(0) = 1, \quad x_2(0) = 1, \\ &\quad \quad \quad u_1^2 + u_2^2 \leq 2, \end{aligned}$$

where p is a real parameter. For $p = 0$, the optimal solution is $u_0 = (-1, -1)$ and $x_0 = (1 - t, 1 - t)$, and the corresponding adjoint variable is $\lambda_0 = (-1, -1)$. The auxiliary problem has the form

$$\begin{aligned} \dot{x}_i &= u_i + s_i, \quad x_i(0) = 1, \quad \dot{\lambda}_i = q_i, \quad \lambda_i(1) = -1, \quad \text{for } i = 1, 2, \\ (\lambda + r)^T(v - u) &\geq 0 \quad \text{whenever } v_1^2 + v_2^2 \leq 2. \end{aligned}$$

The control solving the auxiliary problem is

$$u = \sqrt{2} w/|w|, \quad \text{where } w_i(t) = r_i(t) - 1 + \int_1^t q(s) ds.$$

Hence, the solution of the auxiliary problem is unique and Lipschitz continuous, with respect to $y = (s, r, q)$ around 0, as a function from L^∞ to $W^{1,\infty} \times L^\infty$. By Corollary 2 and for p near zero, there exists a solution (x_p, u_p, λ_p) of the first-order necessary conditions associated with (67), which is unique in a neighborhood of (x_0, u_0, λ_0) and which is a Lipschitz continuous function of p near 0. Using the uniqueness of (x_0, u_0) , it can be shown that (x_p, u_p) is the unique solution of (67). In this example, we use the strong convexity of the constraining set instead of the coercivity condition to ensure Lipschitz continuity.

Remark 8. Lipschitz results for problems with convex cost, linear dynamics, and linear inequality state and control constraints are obtained by Dontchev [13, Chap. 2] using duality theory and the regularity of the optimal control established by Hager [18]. Later, Malanowski [27] studied a problem with a quadratic cost functional, linear inequality state and control constraints, and system dynamics that are linear with respect to the control. A similar problem without state constraints, but with convex control constraints, is considered by Alt in [3] using Robinson's strong regularity condition. In [2] Alt considered a nonlinear problem with inequality control constraints. He obtains an estimate for the optimal control, assuming existence of a solution to the perturbed problem in a neighborhood of the reference point (see Remark 7).

7. Euler's method. Again, let us consider a nonlinear control problem with control constraints, below:

$$(68) \quad \begin{aligned} & \text{minimize} \quad \int_I g(x(t), u(t)) \, dt \\ & \text{subject to} \quad \dot{x}(t) = f(x(t), u(t)) \quad \text{and} \quad u(t) \in U \quad \text{a.e. } t \in I, \\ & \quad \quad \quad x(0) = a, \quad x \in W^{1,\infty}, \quad u \in L^\infty, \end{aligned}$$

where $f: R^{n+m} \rightarrow R^n$, $g: R^{n+m} \rightarrow R$, $U \subset R^m$ is nonempty, closed, and convex, and a is the given starting condition. We assume that there exists a solution (x^*, u^*) to (68) with u^* Riemann integrable, that there exists a closed set $\Delta \subset R^{n+m}$ where both f and g are twice continuously differentiable, and that there exists $\delta > 0$ such that $(x^*(t), u^*(t)) \in \Delta$ and the distance from $(x^*(t), u^*(t))$ to the boundary of Δ is at least δ for every $t \in I$. When we write \dot{x}^* , we mean a function whose values on I coincide with those of $f(x^*, u^*)$.

Let H denote the Hamiltonian defined by

$$H(x, u, \lambda) = g(x, u) + \lambda^T f(x, u),$$

and let $\lambda = \lambda^*$ be the solution of the adjoint equation

$$(69) \quad \dot{\lambda}(t) = -\nabla_x H(x(t), u(t), \lambda(t)) \quad \text{a.e. } t \in I, \lambda(1) = 0,$$

associated with $x = x^*$ and $u = u^*$. By the minimum principle [22, p. 134], we have

$$(70) \quad \nabla_u H(x^*(t), u^*(t), \lambda^*(t))^T (v - u^*(t)) \geq 0 \quad \text{a.e. } t \in I \text{ and for every } v \in U.$$

Given a natural number N , let $h = 1/N$ be the mesh spacing, and let x_i and u_i denote approximations to $x(t)$ and $u(t)$ at $t = t_i = ih$. We consider the Euler discretization of (68) given by

$$(71) \quad \begin{aligned} & \text{minimize} \quad \sum_{i=0}^{N-1} hg(x_i, u_i) \\ & \text{subject to} \quad x_{i+1} = x_i + hf(x_i, u_i) \quad \text{and} \\ & \quad \quad \quad u_i \in U, i = 0, 1, \dots, N-1, x_0 = a. \end{aligned}$$

If (x^h, u^h) denotes a solution to (71), let $\lambda = \lambda^h$ denote the solution of the discrete adjoint equation

$$(72) \quad \lambda_i = \lambda_{i+1} + h \nabla_x H(x_i, u_i, \lambda_{i+1}), \quad i = N-1, N-2, \dots, 0, \lambda_N = 0,$$

associated with $x = x^h$ and $u = u^h$. By the discrete minimum principle [22, p. 280], we have

$$(73) \quad \nabla_u H(x_i^h, u_i^h, \lambda_{i+1}^h)^T (v - u_i^h) \geq 0 \quad \text{for all } v \in U, \quad i = 0, 1, \dots, N-1.$$

To estimate the distance between (x^*, u^*) and (x^h, u^h) , we need a coercivity-type assumption for the discrete problem. Define the following matrices:

$$A(t) = \nabla_{xx} f^*(t), \quad B(t) = \nabla_{xu} f^*(t), \quad Q(t) = \nabla_{xx}^2 H^*(t),$$

$$R(t) = \nabla_{uu}^2 H^*(t), \quad S(t) = \nabla_{xu}^2 H^*(t).$$

Here $f^*(t)$ and $H^*(t)$ stand for $f(x^*(t), u^*(t))$ and $H(x^*(t), u^*(t), \lambda^*(t))$, respectively. Letting A_i, B_i, Q_i, S_i , and R_i denote the corresponding time-varying matrices evaluated at $t = t_i$, we assume that there exists a scalar $\alpha > 0$, α independent of N , such that

$$(74) \quad u^T R_i u \geq \alpha |u|^2, \quad 0 \leq i \leq N-1 \quad \text{whenever } u = v - w \text{ with } v \text{ and } w \in U,$$

$$(75) \quad \sum_{i=0}^{N-1} x_i^T Q_i x_i + u_i^T R_i u_i + 2x_i^T S_i u_i \geq \alpha \sum_{i=0}^{N-1} |u_i|^2$$

whenever $u_i = v_i - w_i$ for some v_i and $w_i \in U$, and

$$(76) \quad x_{i+1} = x_i + h A_i x_i + h B_i u_i, \quad i = 0, 1, \dots, N-1, \quad x_0 = 0.$$

Obviously, the discrete condition (74) holds if there exists $\alpha > 0$ such that

$$u^T R(t) u \geq \alpha |u|^2 \quad \text{for every } t \in I \text{ and for each } u = v - w \text{ with } v \text{ and } w \in U.$$

In Appendix 2, we show that assumption (75) for the discrete problem can be deduced from an analogous assumption for the continuous problem. In analyzing the discrete problem (71), we utilize a discrete L^p norm defined by

$$(\|u\|_{L^p})^p = \sum_{i=0}^{N-1} h |u_i|^p, \quad 1 \leq p < \infty, \quad \text{and} \quad \|u\|_{L^\infty} = \max_{0 \leq i < N} |u_i|.$$

If ϕ and v satisfy the finite difference system

$$\phi_{i+1} = \phi_i + h A_i \phi_i + h v_i, \quad i = 0, 1, \dots, N-1, \quad \phi_0 = 0,$$

then there exists a constant c , independent of h , such that

$$(77) \quad |\phi_j| \leq c \|v\|_{L^1} \leq c \|v\|_{L^2} \quad \text{for each } j = 0, 1, \dots, N.$$

Squaring this inequality, multiplying by h , and summing over j yields

$$\|\phi\|_{L^2}^2 \leq c \|v\|_{L^2}^2.$$

Hence, if the coercivity condition (75) holds relative to the control, then the following joint state-control coercivity condition holds: There exists $\alpha > 0$ such that

$$h \sum_{i=0}^{N-1} x_i^T Q_i x_i + u_i^T R_i u_i + 2x_i^T S_i u_i \geq \alpha (\|x\|_{L^2}^2 + \|u\|_{L^2}^2)$$

whenever $u_i = v_i - w_i$ for some v_i and $w_i \in U$, and

$$x_{i+1} = x_i + h A_i x_i + h B_i u_i, \quad i = 0, 1, \dots, N-1, \quad x_0 = 0.$$

Our convergence result for the discrete problem is expressed in terms of a modulus of smoothness introduced by Sendov and Popov [42]. The local modulus of continuity $\omega(u; t, h)$ of the function u is defined by

$$\omega(u; t, h) = \sup \{|u(a) - u(b)| : a, b \in [t - h/2, t + h/2] \cap I\},$$

while the average modulus of smoothness τ is given by

$$\tau(u; h) = \int_I \omega(u; t, h) dt.$$

In [42, pp. 8–11] it is shown that $\tau(u; h) \rightarrow 0$ as $h \rightarrow 0$ if and only if the bounded function u is Riemann integrable on I ; moreover, $\tau(u; h) = O(h)$ if and only if u has bounded variation on I . The main result in this section is the following theorem.

THEOREM 6. *If u^* is Riemann integrable and the coercivity assumptions (74) and (75) hold, then for all N sufficiently large, there exists a local minimizer (x^h, u^h) of (71) such that*

$$\max_{0 \leq i \leq N-1} |u^*(t_i) - u_i^h| = O(h + \tau(u^*; h)),$$

$$\max_{0 \leq i \leq N} |x^*(t_i) - x_i^h| = O(h + \tau(u^*; h)),$$

$$\max_{0 \leq i \leq N} |\lambda^*(t_i) - \lambda_i^h| = O(h + \tau(u^*; h)),$$

$$\max_{0 \leq i \leq N-1} \left| \dot{x}^*(t_i) - \frac{x_{i+1}^h - x_i^h}{h} \right| = O(h + \tau(u^*; h)).$$

Hence, if u^* has bounded variation, then each of these error estimates is of order h .

Proof. We apply Corollary 1 to the necessary conditions associated with the discrete problem (71). The parameter p of Corollary 1 is identified with the mesh spacing h the set Ω_p consists of discrete triples (x, u, λ) , where $u_i \in U$ for each i . Component i , $0 \leq i \leq N-1$, of the operators T_p and F_p , denoted T_i^h and F_i^h , respectively, is the following:

$$T_i^h(x, u, \lambda) = \begin{bmatrix} \nabla_x H(x_i, u_i, \lambda_{i+1}) + (\lambda_{i+1} - \lambda_i)/h \\ \nabla_u H(x_i, u_i, \lambda_{i+1}) \\ f(x_i, u_i) - (x_{i+1} - x_i)/h \end{bmatrix} \quad \text{and} \quad F_i^h(x, u, \lambda) = \begin{bmatrix} 0 \\ \partial U(u_i) \\ 0 \end{bmatrix}.$$

Given $z = (x, u, \lambda)$ in the discrete space Z_p associated with Ω_p , we use the L^∞ norm for each of the three components x , u , and λ of z . In the discrete space Y_p associated with the range of T_p , we use the L^1 norm for the first and last component, and the L^∞ norm for the middle component. That is, if $y = (a, b, c) \in Y_p$, then

$$\|y\|_p = \|a\|_{L^1} + \|b\|_{L^\infty} + \|c\|_{L^1}.$$

The point z_p of Corollary 1 is given by $z_p = (x^I, u^I, \lambda^I)$, where

$$x_i^I = x^*(t_i), \quad u_i^I = u^*(t_i), \quad \lambda_i^I = \lambda^*(t_i).$$

Also, in Corollary 1, component i of the point y_p , denoted y_i^h , is the triple

$$y_i^h = \begin{bmatrix} 0 \\ \nabla_u H(x_i^I, u_i^I, \lambda_i^I) \\ 0 \end{bmatrix}.$$

Observe that with this choice for y_p , we have $y_p \in F(z_p)$. We define a linear operator M^h that acts on a discrete triple (x, u, λ) to produce a vector whose i th component is

$$M_i^h(x, u, \lambda) = \begin{bmatrix} A_i^T \lambda_{i+1} + Q_i x_i + S_i u_i + (\lambda_{i+1} - \lambda_i)/h \\ R_i u_i + S_i^T x_i + B_i^T \lambda_{i+1} \\ A_i x_i + B_i u_i - (x_{i+1} - x_i)/h \end{bmatrix}.$$

Taking $L_p(z) = M^h(x, u, \lambda) - M^h(x^l, u^l, \lambda^l)$, observe that $L_p(z_p) = 0$.

It can be verified that under the smoothness assumptions and for ρ smaller than δ , we have $D_\rho(h) \rightarrow 0$ as $h \rightarrow 0$. Now consider the term

$$(78) \quad (T_p(z_p) - y_p)_i = \begin{bmatrix} \nabla_x H(x_i^l, u_i^l, \lambda_{i+1}^l) + (\lambda_{i+1}^l - \lambda_i^l)/h \\ \nabla_u H(x_i^l, u_i^l, \lambda_{i+1}^l) - \nabla_u H(x_i^l, u_i^l, \lambda_i^l) \\ f(x_i^l, u_i^l) - (x_{i+1}^l - x_i^l)/h \end{bmatrix}.$$

The middle component of this vector is $O(h)$ since λ^* is Lipschitz continuous. Since the analysis of the first and last component in (78) is similar, we only focus on the last component

$$(79) \quad \left| f(x_i^l, u_i^l) - \frac{x_{i+1}^l - x_i^l}{h} \right| \leq \frac{1}{h} \int_{t_i}^{t_{i+1}} |f(x_i^l, u_i^l) - \dot{x}^*(t)| dt \\ \leq \frac{1}{h} \int_{t_i}^{t_{i+1}} |f(x_i^l, u_i^l) - f(x^*(t), u^*(t))| dt \leq \frac{c}{h} \int_{t_i}^{t_{i+1}} [h + \omega(u^*; t, 2h)] dt,$$

where c denotes a generic constant, independent of h . Multiplying (79) by h , summing over i , and exploiting the inequality $\tau(u; kh) \leq k\tau(u; h)$ for each natural number k (see [42, p. 11]), it follows that

$$\|T_p(z_p) - y_p\|_p = O(h + \tau(u^*; h)).$$

Next, we must analyze the auxiliary problem and establish the existence of a constant γ satisfying (9). The analysis essentially parallels that of [20] except that continuous norms are replaced by their discrete analogues. We must examine how the solution to the following system depends on the perturbations q_i , r_i , and s_i :

$$(80) \quad \begin{aligned} A_i^T \lambda_{i+1} + Q_i x_i + S_i u_i + \frac{\lambda_{i+1} - \lambda_i}{h} + q_i &= 0, & \lambda_N &= 0, \\ (R_i u_i + S_i^T x_i + B_i^T \lambda_{i+1} + r_i)(v - u_i) &\geq 0 & \text{for every } v &\in U, \\ A_i x_i + B_i u_i - \frac{x_{i+1} - x_i}{h} + s_i &= 0, & x_0 &= a, \end{aligned}$$

$i = 0, 1, \dots, N-1$. Note that system (80) constitutes the first-order necessary conditions (see [22, p. 280]) associated with the following quadratic program:

$$(81) \quad \begin{aligned} \text{minimize} \quad & h \sum_{i=0}^{N-1} \frac{1}{2} x_i^T Q_i x_i + \frac{1}{2} u_i^T R_i u_i + x_i^T S_i u_i + q_i^T x_i + r_i^T u_i \\ \text{subject to} \quad & x_{i+1} = x_i + hA_i x_i + hB_i u_i + hs_i \quad \text{and} \\ & u_i \in U, \quad 0 \leq i \leq N-1, \quad x_0 = a. \end{aligned}$$

By Lemma 4 and the discussion that follows it, there is a one-to-one correspondence between a solution to (81) and a solution to (80) when (75) holds.

Now consider the perturbations (q^i, r^i, s^i) for $i = 1$ and 2 . Let (x^i, u^i, λ^i) denote the associated solutions to (80). Referring to Lemma 4 (see [20, § 2] for more details), we have

$$\|u^1 - u^2\|_{L^2} \leq c(\|q^1 - q^2\|_{L^1} + \|r^1 - r^2\|_{L^2} + \|s^1 - s^2\|_{L^1}),$$

where c is a constant independent of h . Utilizing (77), we also conclude that

$$(82) \quad \|x^1 - x^2\|_{L^\infty} + \|\lambda^1 - \lambda^2\|_{L^\infty} \leq c(\|q^1 - q^2\|_{L^1} + \|r^1 - r^2\|_{L^2} + \|s^1 - s^2\|_{L^1}).$$

Finally, by (74) and Lemma 4, we have

$$\|u^1 - u^2\|_{L^\infty} \leq c(\|r^1 - r^2\|_{L^\infty} + \|x^1 - x^2\|_{L^\infty} + \|\lambda^1 - \lambda^2\|_{L^\infty}).$$

Combining this with (82) yields

$$\begin{aligned} & \|x^1 - x^2\|_{L^\infty} + \|u^1 - u^2\|_{L^\infty} + \|\lambda^1 - \lambda^2\|_{L^\infty} \\ & \leq c(\|q^1 - q^2\|_{L^1} + \|r^1 - r^2\|_{L^\infty} + \|s^1 - s^2\|_{L^1}). \end{aligned}$$

Hence, there exists a constant γ such that (9) holds with $\sigma = \infty$.

By Corollary 1, there exists a solution to the discrete necessary conditions (72) and (73) associated with (71) that satisfies the first three estimates of Theorem 6. The discrete and continuous state equations, along with the previously established error estimates, imply that

$$\left| x^*(t_i) - \frac{x_{i+1}^h - x_i^h}{h} \right| = |f(x^*(t_i), u^*(t_i)) - f(x_i^h, u_i^h)| = O(h + \tau(u^*; h)),$$

which gives the last estimate of Theorem 6. The fact that x^h and u^h are local minimizers for (71) follows from Corollary 6 in Appendix 1, the coercivity condition (75), and the fact that coercivity condition (75) is preserved after small perturbations in Q_i , R_i , S_i , A_i , and B_i . \square

Remark 9. Note that the coercivity assumptions (74) and (75) do not necessarily imply that an optimal control is either unique or continuous. For example, if $g(x, u) = (u^2 - 1)^2$, $f = 0$, and $U = R^1$, then for each measurable set $M \subset 1$, the function defined by

$$u(t) = 1 \quad \text{for } t \in M \quad \text{and} \quad u(t) = -1 \quad \text{for } t \notin M$$

is an optimal control that satisfies (74) and (75).

Remark 10. Results most closely related to Theorem 6 include the papers of Budak, Berkovich, and Solov'eva [8] and Cullum [11] in which convergence of the optimal value associated with discrete approximations to state and control constrained problems is established. Mordukhovich [31] shows that the discrete optimal cost converges to the true optimal cost if and only if a relaxation of the control problem is stable. Estimates for the error in the optimal control associated with higher-order discretizations of unconstrained nonlinear problems are derived by Hager [17]. Dontchev [12] obtains an error estimate for Euler's approximation applied to an optimal control problem with convex cost, linear system dynamics, and linear inequality state and control constraints.

Appendix 1: Sufficient optimality conditions. We begin by establishing the sufficient optimality result needed for Corollary 4. Let us consider the following optimization problem:

$$(83) \quad \begin{aligned} & \text{minimize} && C(z) \\ & \text{subject to} && g(z) \in K_g, \quad h(z) \in K_h, \end{aligned}$$

where $g: Z \rightarrow W_g$ and $h: Z \rightarrow W_h$, W_g and W_h are Banach spaces, and K_g and K_h are closed, convex cones with vertices at the origin of their respective spaces. The Lagrangian H associated with (83) is given by

$$H(z, \mu, \nu) = C(z) - \langle \mu, g(z) \rangle - \langle \nu, h(z) \rangle,$$

where $\mu \in W_g^*$ and $\nu \in W_h^*$. Letting z^* be a point that is feasible in (83), we assume that C , g , and h are twice Fréchet differentiable at z^* . The first-order necessary conditions associated with (83) have the following form: There exists $\mu \in K_g^+$ and $\nu \in K_h^+$ such that

$$(84) \quad \nabla_z H(z^*, \mu, \nu) = 0, \quad \langle \mu, g(z^*) \rangle = 0, \quad \langle \nu, h(z^*) \rangle = 0.$$

Although the following lemma makes the same surjectivity assumption that appears in Corollary 4, this assumption can be replaced by any condition that ensures regularity of the linearized system (see Robinson [35] or Maurer and Zowe [30]).

LEMMA 8. *Suppose that z^* is feasible in (83), $\mu \in K_g^+$, $\nu \in \text{int } K_h^+$, the first-order necessary conditions (84) hold, the operator*

$$\begin{bmatrix} h'(z^*) \\ g'(z^*) \end{bmatrix}$$

is surjective, and there exists $\alpha > 0$ such that

$$\begin{aligned} &\langle \nabla_{zz}^2 H(z^*, \mu, \nu)(z - z^*), z - z^* \rangle \geq \alpha \|z - z^*\|^2 \\ &\text{whenever } g(z^*) + g'(z^*)(z - z^*) \in K_g \quad \text{and} \quad h'(z^*)(z - z^*) = 0. \end{aligned}$$

Then z^ is a strict local minimizer for (83).*

In comparing this result to Maurer and Zowe's classic sufficient optimality result [30], observe that the constraint $h'(z^*)(z - z^*) = 0$ in the coercivity condition above corresponds to a constraint of the form $h'(z^*)(z - z^*) \in K_h$ in [30]. In this respect, the coercivity condition of Lemma 8 is weaker than that of [30]. On the other hand, Lemma 8 assumes that $\nu \in \text{int } K_h^+$, while [30] only assumes that $\nu \in K_h^+$.

Proof. Throughout this proof, we let ε denote a generic positive constant that can be made arbitrarily small for z sufficiently close to z^* , we let α denote a generic positive constant that is uniformly bounded away from zero for z near z^* , and we let β denote a generic constant that is uniformly bounded from above for z near z^* . Expanding $H(z, \mu, \nu)$ in a Taylor series about $z = z^*$, we have

$$\begin{aligned} H(z, \mu, \nu) &= H(z^*, \mu, \nu) + \nabla_z H(z^*, \mu, \nu)(z - z^*) \\ &\quad + \frac{1}{2} \nabla_{zz}^2 H(z^*, \mu, \nu)(z - z^*, z - z^*) + R(z), \end{aligned}$$

where $R(z) \leq \varepsilon \|z - z^*\|^2$. By the first-order necessary conditions, $H(z^*, \mu, \nu) = C(z^*)$ and $\nabla_z H(z^*, \mu, \nu) = 0$. Hence, it follows that

$$C(z) = C(z^*) + M(z) + R(z),$$

$$\text{where } M(z) = \langle \mu, g(z) \rangle + \langle \nu, h(z) \rangle + \frac{1}{2} \nabla_{zz}^2 H(z^*, \mu, \nu)(z - z^*, z - z^*).$$

If z is feasible in (83), then since $\nu \in \text{int } K_h^+$, we have $\langle \nu, h(z) \rangle \geq \alpha \|h(z)\|$. Thus we have

$$\langle \mu, g(z) \rangle + \langle \nu, h(z) \rangle \geq \alpha \|h(z)\|.$$

By the complementary slackness condition, $h(z^*) = 0$, and by the differentiability assumption,

$$h(z) = h'(z^*)(z - z^*) + o(\|z - z^*\|) \quad \text{and} \quad g(z) = g(z^*) + g'(z^*)(z - z^*) + o(\|z - z^*\|).$$

Referring to [35, Thm. 1], the surjectivity assumption implies that for each z near z^* , there exists an associated $y \in Z$ such that $h'(z^*)(y - z^*) = 0$, $g(z^*) + g'(z^*)(y - z^*) \in K_g$, and

$$\|y - z\| \leq \beta \|h'(z^*)(z - z^*)\| + \beta \inf \{\|g(z^*) + g'(z^*)(z - z^*) - k\| : k \in K_g\}.$$

This bound, combined with the Taylor expansions of g and h , implies that for each z near z^* , with z feasible in (83), there exists an associated $y \in Z$ such that $h'(z^*)(y - z^*) = 0$, $g(z^*) + g'(z^*)(y - z^*) \in K_g$, and $\|y - z\| \leq \beta \|h(z)\| + \varepsilon \|z - z^*\|$. Applying the triangle inequality yields

$$\|y - z\| \leq \beta \|h(z)\| + \varepsilon \|y - z^*\|.$$

By the coercivity assumption,

$$\nabla_{zz}^2 H(z^*, \mu, \nu)(z - z^*, z - z^*) \geq \alpha \|y - z^*\|^2 - \beta \|y - z^*\| \|y - z\| - \beta \|y - z\|^2.$$

These inequalities, along with the relation $h(z^*) = 0$, imply that for z near z^* with z feasible in (83), we have

$$M(z) \geq \alpha \|h(z)\| + \alpha \|y - z^*\|^2.$$

Recall that the remainder term R has the bound $R(z) \leq \varepsilon \|z - z^*\|^2$. By the triangle inequality,

$$\|z - z^*\| \leq \|z - y\| + \|y - z^*\| \leq (1 + \varepsilon) \|y - z^*\| + \beta \|h(z)\|,$$

from which it follows that

$$\|z - z^*\|^2 \leq \beta \|y - z^*\|^2 + \beta \|h(z)\|^2.$$

Hence, for z near z^* with z feasible in (83), we have

$$\begin{aligned} C(z) - C(z^*) &= M(z) + R(z) \geq \alpha \|h(z)\| + \alpha \|y - z^*\|^2 - \varepsilon \|z - z^*\|^2 \\ &\geq \alpha \|h(z)\| + \alpha \|z - z^*\|^2 - \beta \|h(z)\|^2 \geq \alpha \|z - z^*\|^2, \end{aligned}$$

which completes the proof. \square

Next, we obtain sufficient optimality conditions that are applicable to optimal control problems. A number of relevant sufficient optimality conditions have appeared in the literature; for example, see Ioffe [21] and, in particular, the results of Maurer [29]. Although the basic strategy for obtaining sufficient optimality results in the optimal control setting is developed nicely by Maurer in [29], the precise results that we need in §§ 6 and 7 are not stated in his paper. For completeness, we give a brief, self-contained treatment of the results needed in our paper. We begin with the abstract problem

$$(85) \quad \begin{aligned} &\text{minimize} && C(z) \\ &\text{subject to} && z \in \Lambda, \end{aligned}$$

where Λ is a subset of a normed vector space Z , and C is a real-valued function. As in [29], we assume that there are two different norms, denoted by $\|\cdot\|$ and $\|\cdot\|$, associated with Z .

LEMMA 9. *Suppose that z^* satisfies the constraints of (85) and that there exists a functional M and a scalar $\alpha > 0$ with the following property:*

$$(86) \quad M(z) \geq \alpha \|z - z^*\|^2 \quad \text{for each } z \in \Lambda \text{ with } \|z - z^*\| \text{ sufficiently small}$$

and

$$(87) \quad \frac{C(z) - C(z^*) - M(z)}{\|z - z^*\|^2} \rightarrow 0 \quad \text{as } \|z - z^*\| \rightarrow 0 \quad \text{with } z \in \Lambda.$$

Then z^* is a strict local minimizer for (85).

Proof. By the hypotheses above, we have

$$C(z) - C(z^*) \geq \alpha \|z - z^*\|^2 + o(\|z - z^*\|^2)$$

as $\|z - z^*\| \rightarrow 0$ with $z \in \Lambda$, which implies that z^* is a strict local minimizer for (85). \square

In the application of Lemma 9, the following observation is helpful.

LEMMA 10. Suppose that there exists a scalar $\alpha > 0$, a bilinear form b that is bounded relative to the norm $\|\cdot\|$, and a set T such that

$$b(z - z^*, z - z^*) \geq \alpha \|z - z^*\|^2 \quad \text{for every } z \in T.$$

If for each $z \in \Lambda$, there exists $y \in T$ such that $\|z - y\| = o(\|z - z^*\|)$, then for each $\beta < \alpha$, we have

$$b(z - z^*, z - z^*) \geq \beta \|z - z^*\|^2 \quad \text{for all } z \in \Lambda \text{ with } \|z - z^*\| \text{ sufficiently small.}$$

Proof. Given $z \in \Lambda$, let $y \in T$ be the hypothesized point for which $\|z - y\| = o(\|z - z^*\|)$. Since the bilinear form b is bounded, there exists a constant c such that

$$b(z - z^*, z - z^*) \geq \alpha \|y - z^*\|^2 - c \|z - y\|^2 - c \|z - y\| \|y - z^*\|.$$

The inequality

$$\|y - z^*\| \geq \|z - z^*\| - \|z - y\| = \|z - z^*\| - o(\|z - z^*\|)$$

completes the proof. \square

We now apply Lemmas 9 and 10 to optimal control problems. Note that in the following result, we neither assume an interior point nor controllability.

COROLLARY 5. Suppose that x^* and u^* are feasible for the optimal control problem (68), that f and g satisfy the differentiability conditions given below (68), that $\lambda = \lambda^*$ is the solution to the adjoint equation (69) associated with $x = x^*$ and $u = u^*$, and that the minimum principle (70) holds. If there exists $\alpha > 0$ such that

$$(88) \quad \int_I (x(t)^T Q(t)x(t) + u(t)^T R(t)u(t) + 2x(t)^T S(t)u(t)) dt \geq \alpha \int_I |u(t)|^2 dt$$

whenever $x \in W^{1,2}$, $x(0) = 0$, $u \in L^2$, $\dot{x} = Ax + Bu$, $u = v - u^*$ for some $v \in L^2$ with $v(t) \in U$ for almost every $t \in I$, then u^* is a strict local minimizer for (68).

Proof. We apply Lemmas 9 and 10 with the following identifications: The z of Lemma 9 is the pair (x, u) , the space Z is $W^{1,\infty} \times L^\infty$, the norm $\|\cdot\|$ associated with Z is the L^2 inner product norm, $C(z)$ is the integral cost function in (68), and Λ consists of those (x, u) in a convex neighborhood of (x^*, u^*) that satisfy the constraints

$$F(z) = 0, \quad \text{where } F(z) = f(x, u) - \dot{x}, \quad u(t) \in U \quad \text{a.e. } t \in I, \quad \text{and } x(0) = a.$$

The functional M is defined by

$$M(z) = \langle \nabla_u H(x^*, u^*, \lambda^*), u - u^* \rangle + b(z - z^*, z - z^*), \quad z = (x, u),$$

where

$$b(\delta z, \delta z) = \int_I (\delta x(t)^T Q(t) \delta x(t) + \delta u(t)^T R(t) \delta u(t) + 2 \delta x(t)^T S(t) \delta u(t)) dt,$$

$$\delta z = (\delta x, \delta u).$$

The set T of Lemma 10 is given by

$$T = \{z = (x, u) \in Z: F'(z^*)(z - z^*) = 0, x(0) = a, u(t) \in U \text{ a.e. } t \in I\}.$$

To verify identity (87), first note that for $z \in \Lambda$, we have

$$C(z) = \int_I g(z) + F(z)^T \lambda^* dt.$$

Hence, after an integration by parts and a Taylor expansion, we obtain (87). By the minimum principle (70), $M(z) \geq b(z - z^*, z - z^*)$. By the coercivity condition (88) and the fact that coercivity with respect to the control implies coercivity with respect to the state (see [20]), we have

$$b(z - z^*, z - z^*) \geq \alpha \|z - z^*\|^2 \quad \text{for some } \alpha > 0.$$

Thus (86) follows from Lemma 10 if, for each $z = (x, u) \in \Lambda$, we can establish the existence of $y \in T$ with $\|z - y\| = o(\|z - z^*\|)$. We construct y in the following way: Let w be the solution to

$$(89) \quad \dot{w} = \nabla_x f(x^*, u^*)w - F'(z^*)(z - z^*), \quad w(0) = 0,$$

and define $y = (w + x, u)$. Observe that $y \in T$. Also, by (89) we have

$$\|w\|_{L^2} \leq c \|F'(z^*)(z - z^*)\|_{L^2},$$

where c is a generic constant. From the relation

$$\|F'(z^*)(z - z^*)\|_{L^2} = \|F(z) - F(z^*) - F'(z^*)(z - z^*)\|_{L^2} = o(\|z - z^*\|_{L^2}),$$

we conclude that $\|z - y\|_{L^2} = \|w\|_{L^2} = o(\|z - z^*\|_{L^2})$, which completes the proof. \square

Now let us consider the finite-dimensional optimization problem

$$(90) \quad \begin{aligned} &\text{minimize} && C(x, u) \\ &\text{subject to} && F(x, u) = 0, \quad u \in \Omega \subset R^m, \quad x \in R^n, \end{aligned}$$

where Ω is convex and F maps R^{m+n} to R^n . Let z denote the pair (x, u) , and for $\lambda \in R^n$, let H be the Lagrangian defined by

$$H(z, \lambda) = C(z) + \lambda^T F(z).$$

COROLLARY 6. *Suppose that x^* and u^* are feasible for (90), that F and C are twice differentiable at $z^* = (x^*, u^*)$, and that $\nabla_x F(x^*, u^*)$ is nonsingular. If there exists a multiplier $\lambda^* \in R^n$ such that*

$$\nabla_x H(x^*, u^*, \lambda^*) = 0 \quad \text{and} \quad \nabla_u H(x^*, u^*, \lambda^*)^T (u - u^*) \geq 0 \quad \text{for every } u \in \Omega$$

and

$$(z - z^*)^T \nabla_{zz}^2 H(z^*, \lambda^*) (z - z^*) \geq \alpha |z - z^*|^2$$

whenever $F'(z^*)(z - z^*) = 0$ for some $z = (x, u)$ with $u \in \Omega$, then z^* is a local minimizer for (90).

Proof. We apply Lemma 9 with

$$M(z) = \nabla_z H(z^*, \lambda^*) (z - z^*) + \frac{1}{2} (z - z^*)^T \nabla_{zz}^2 H(z^*, \lambda^*) (z - z^*).$$

The set T of Lemma 10 is given by

$$T = \{z = (x, u) \in R^{m+n}: F'(z^*)(z - z^*) = 0, u \in \Omega\}.$$

The y of Lemma 10 is constructed in the following way: Given $z = (x, u)$ with $u \in \Omega$, $y = (x + w, u)$, where w is the solution to

$$\nabla_x F(x^*, u^*) w = -F'(z^*)(z - z^*).$$

Observe that $y \in T$. If $F(z) = 0$, then

$$|F'(z^*)(z - z^*)| = |F(z) - F(z^*) - F'(z^*)(z - z^*)| = o(|z - z^*|).$$

Since $\nabla_x F(x^*, u^*)$ is nonsingular, we have

$$|z - y| = |w| \leq c|F'(z^*)(z - z^*)| = o(|z - z^*|). \quad \square$$

Appendix 2: The coercivity condition. Here we show that the discrete coercivity condition (75) of § 7 can be deduced from an analogous continuous condition.

LEMMA 11. *Suppose that the matrices A , B , Q , R , and S of § 7 are continuous and that there exists $\beta > 0$ such that*

$$(91) \quad \int_I (x(t)^T Q(t)x(t) + u(t)^T R(t)u(t) + 2x(t)^T S(t)u(t)) dt \geq \beta \int_I |u(t)|^2 dt$$

whenever $x \in W^{1,2}$, $x(0) = 0$, $u \in L^2$, $\dot{x} = Ax + Bu$, $u = v - w$ for some v and $w \in L^2$ with $v(t)$ and $w(t) \in U$ for almost every $t \in I$. Then there exists $\alpha > 0$ satisfying the discrete coercivity condition (75).

Proof. Given sequences $\{x_i\}$ and $\{u_i\}$ that satisfy the linear equation (76) where $u_i = v_i - w_i$ for some v_i and $w_i \in U$, let u^h denote the piecewise constant extension of the u_i defined by

$$u^h(t) = u_i, \quad t_i \leq t < t_{i+1}, \quad i = 0, 1, \dots, N-1,$$

and let x^h be the solution of

$$\dot{x}^h = Ax^h + Bu^h, \quad x^h(0) = 0.$$

Define $y_i = x^h(t_i)$ and let y^h be the piecewise constant extension of the y_i . Since u^h is piecewise constant,

$$\int_I |u^h(t)|^2 dt = h \sum_{i=0}^{N-1} |u_i|^2.$$

We will show that, for $x = x^h$ and $u = u^h$,

$$(92) \quad |\text{left side of (91)} - \text{left side of (75)}| \leq h\epsilon^h \sum_{i=0}^{N-1} |u_i|^2,$$

where ϵ^h denotes a generic constant that tends to zero as h tends to zero. Hence, (75) follows from (91) when h is sufficiently small.

Let us begin with the quadratic control terms in (75) and (91). Since u^h is equal to u_i on the interval $[t_i, t_{i+1}]$, it follows that

$$h \sum_{i=0}^{N-1} u_i^T R_i u_i - \int_I u^h(t)^T R(t) u^h(t) dt = h \sum_{i=0}^{N-1} u_i^T \delta R_i u_i,$$

where

$$(93) \quad \delta R_i = R_i - \frac{1}{h} \int_{t_i}^{t_{i+1}} R(t) dt.$$

Since $R(t)$ is continuous in t , (93) approaches zero, uniformly in i , as $N \rightarrow \infty$. Hence, we have

$$\left| h \sum_{i=0}^{N-1} u_i^T R_i u_i - \int_I u^h(t)^T R(t) u^h(t) dt \right| \leq h \varepsilon^h \sum_{i=0}^{N-1} |u_i|^2.$$

Now let us consider the quadratic state terms in (75) and (91). As with the quadratic control term, we have

$$(94) \quad \left| h \sum_{i=0}^{N-1} y_i^T Q_i y_i - \int_I y^h(t)^T Q(t) y^h(t) dt \right| \leq h \varepsilon^h \sum_{i=0}^{N-1} |y_i|^2.$$

From the differential equation satisfied by x^h , we have

$$(95) \quad \|y^h\|_{L^2}^2 \leq \|y^h\|_{L^\infty}^2 \leq \|x^h\|_{L^\infty}^2 \leq c \|u^h\|_{L^2}^2 = ch \sum_{i=0}^{N-1} |u_i|^2,$$

where c denotes a generic constant that is independent of h for h sufficiently small. Combining (94) and (95) yields

$$(96) \quad \left| h \sum_{i=0}^{N-1} y_i^T Q_i y_i - \int_I y^h(t)^T Q(t) y^h(t) dt \right| \leq h \varepsilon^h \sum_{i=0}^{N-1} |u_i|^2.$$

Since y^h is the piecewise constant extension of x^h , it follows from the equation for x^h that

$$\|y^h - x^h\|_{L^2} \leq h \|\dot{x}^h\|_{L^2} \leq ch \|u^h\|_{L^2}.$$

This estimate, along with (95), implies that

$$(97) \quad \left| \int_I y^h(t)^T Q(t) y^h(t) - x^h(t)^T Q(t) x^h(t) dt \right| \leq ch \|u^h\|_{L^2}^2 = h \varepsilon^h \sum_{i=0}^{N-1} |u_i|^2.$$

Finally, let us consider the difference

$$\sum_{i=0}^{N-1} y_i^T Q_i y_i - x_i^T Q_i x_i.$$

Integrating the differential equation for x^h over the interval $[t_i, t_{i+1}]$ gives

$$(98) \quad y_{i+1} = y_i + h A_i y_i + h B_i u_i + e_i,$$

where

$$e_i = -h \delta B_i u_i - h \delta A_i y_i + \int_{t_i}^{t_{i+1}} A(t) (x^h(t) - y_i) dt.$$

The factors δA_i and δB_i are defined by

$$\delta A_i = A_i - \frac{1}{h} \int_{t_i}^{t_{i+1}} A(t) dt \quad \text{and} \quad \delta B_i = B_i - \frac{1}{h} \int_{t_i}^{t_{i+1}} B(t) dt.$$

Subtracting the finite difference equation (76) from (98) gives

$$|y_j - x_j| \leq c \sum_{i=0}^{N-1} |e_i|.$$

From the definition of e_i , we have

$$|e_i| \leq h \varepsilon^h (|u_i| + |y_i|).$$

It follows that

$$|y_j - x_j| \leq h \varepsilon^h \sum_{i=0}^{N-1} |u_i| + |y_i| \leq \varepsilon^h \|u^h\|_{L^2}.$$

Summing over j yields

$$h \sum_{j=0}^{N-1} |y_j - x_j|^2 \leq \varepsilon^h \|u^h\|_{L^2}^2.$$

Hence, we have

$$(99) \quad \left| h \sum_{i=0}^{N-1} y_i^T Q_i y_i - x_i^T Q_i x_i \right| \leq \varepsilon^h \|u^h\|_{L^2}^2.$$

The triangle inequality, along with (96), (97), and (99), gives

$$\left| h \sum_{i=0}^{N-1} x_i^T Q_i x_i - \int_I x^h(t)^T Q(t) x^h(t) dt \right| \leq h \varepsilon^h \sum_{i=0}^{N-1} |u_i|^2.$$

Since the cross-product term $x_i^T S_i u_i$ can be analyzed in a similar manner, the proof of (92) is complete. \square

REFERENCES

- [1] W. ALT, *Stability of solutions for a class of nonlinear cone constrained optimization problems, Part 1: Basic theory*, Numer. Funct. Anal. Optim., 10 (1989), pp. 1053–1064.
- [2] ———, *Stability of solutions to control constrained nonlinear optimal control problems*, Appl. Math. Optim., 21 (1990), pp. 53–68.
- [3] ———, *Parametric optimization with applications to optimal control and sequential quadratic programming*, Bayreuth. Math. Schr., 35 (1991), pp. 1–37.
- [4] J.-P. AUBIN, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.
- [5] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.
- [6] J.-P. AUBIN AND H. FRANKOWSKA, *On inverse function theorems for set-valued maps*, J. Math. Pures Appl., 66 (1987), pp. 71–89.
- [7] ———, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
- [8] B. M. BUDAK, E. M. BERKOVICH, AND E. N. SOLO'EVA, *The convergence of difference approximations in optimal control problems*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 522–547.
- [9] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [10] R. COMINETTI, *Metric regularity, tangent sets, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.
- [11] J. CULLUM, *Finite-dimensional approximations of state-constrained continuous optimal control problems*, SIAM J. Control, 10 (1972), pp. 649–670.
- [12] A. L. DONTCHEV, *Error estimates for a discrete approximation to constrained control problems*, SIAM J. Numer. Anal., 13 (1981), pp. 500–514.
- [13] ———, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci., Vol. 52, Springer-Verlag, New York, 1983.
- [14] ———, *On the admissible controls of constrained linear systems*, C. R. Acad. Bulgare Sci., 42 (1989), pp. 33–36.
- [15] J. C. DUNN AND T. TIAN, *Variants of the Kuhn–Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.
- [16] H. FRANKOWSKA, *Some inverse mapping theorems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 183–234.
- [17] W. W. HAGER, *Rate of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–471.
- [18] ———, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.
- [19] ———, *Approximations to the multiplier method*, SIAM J. Numer. Anal., 22 (1985), pp. 16–46.
- [20] ———, *Multiplier methods for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061–1080.

- [21] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum. 3: Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.
- [22] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
- [23] K. ITO AND K. KUNISCH, *Sensitivity analysis of solutions to optimization problems in Hilbert spaces with applications to optimal control and estimation*, J. Differential Equations, 99 (1992), pp. 1–40.
- [24] A. J. KING AND R. T. ROCKAFELLAR, *Sensitivity analysis for nonsmooth generalized equations*, Math. Programming, 55 (1992), pp. 192–212.
- [25] S. KURCYUSZ, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Optim. Theory Appl., 20 (1976), pp. 81–110.
- [26] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, S. K. Mitter, trans., Springer-Verlag, Berlin, New York, 1971.
- [27] K. MALANOWSKI, *On stability of solutions to constrained optimal control problems for systems with control appearing linearly*, Arch. Automat. Telemek., 33 (1988), pp. 483–497.
- [28] ———, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimizations problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.
- [29] H. MAURER, *First and second-order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.
- [30] H. MAURER AND J. ZOWE, *First- and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
- [31] B. S. MORDUKHOVICH, *On difference approximations of optimal control systems*, Appl. Math. Mech., 42 (1978), pp. 452–461.
- [32] ———, *Sensitivity analysis in nonsmooth optimization*, in Theoretical Aspects of Industrial Design, SIAM, Philadelphia, PA, 1992, pp. 32–46.
- [33] J.-P. PENOT, *Metric regularity, openness, and Lipschitz multifunctions*, Nonlinear Anal. Theory Methods Appl., 13 (1989), pp. 629–643.
- [34] S. M. ROBINSON, *An inverse-function theorem for a class of multivalued functions*, Proc. Amer. Math. Soc., 41 (1973), pp. 211–218.
- [35] ———, *Stability theory for systems of inequalities, Part I: Linear systems*, SIAM J. Numer. Anal., 12 (1976), pp. 754–772.
- [36] ———, *Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
- [37] ———, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
- [38] ———, *An implicit-function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.
- [39] R. T. ROCKAFELLAR, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867–885.
- [40] ———, *Lipschitzian stability in optimization: The role of nonsmooth analysis*, in Nondifferentiable Optimization: Motivation and Applications, V. F. Demyanov and D. Pallaschke, eds., Lecture Notes in Economics and Mathematical Systems, Vol. 255, Springer-Verlag, New York, 1985, pp. 55–73.
- [41] ———, *Proto-differentiability of set-valued mappings and its application in optimization*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 449–482.
- [42] B. SENDOV AND V. A. POPOV, *The Averaged Moduli of Smoothness*, John Wiley, New York, 1988.
- [43] C. URSESCU, *Multifunctions with closed convex graph*, Czechoslovak Math. J., 25 (1975), pp. 438–441.

PURSUIT-EVASION PROBLEMS AND VISCOSITY SOLUTIONS OF ISAACS EQUATIONS*

PIERPAOLO SORAVIA†

Abstract. The Dirichlet problem for first-order Hamilton–Jacobi equations arising in differential games of pursuit and evasion is studied. Local and global sub- and superoptimality principles are stated for, respectively, viscosity sub- and supersolutions. These results are applied to obtain a general existence theorem and to prove the existence of the value of the game. The main application concerns the problem of stability (terminability) of a dynamical system with two competitive controls and the opposite one of evadability from a general closed set. The approach used in this paper allows Lyapunov functions satisfying the usual condition in the weak sense of viscosity solutions.

Key words. viscosity solutions, comparison theorems, dynamical systems, stability, differential games

AMS(MOS) subject classifications. 35B37, 49L25, 90D25

Introduction. In a previous work of Bardi and the author [BS1], we studied the Dirichlet problem for a Hamilton–Jacobi equation

$$(0.1) \quad \begin{aligned} H(x, DU) &= 0 && \text{in } \Omega \setminus \mathcal{T}, \\ U &= g && \text{on } \partial\mathcal{T}, \\ U(x) &\rightarrow +\infty && \text{as } x \rightarrow x_0 \in \partial\Omega, \end{aligned}$$

where the Hamiltonian can be written in the following way:

$$(0.2) \quad H(x, p) := \inf_{b \in B} \sup_{a \in A} \{-f(x, a, b) \cdot p - h(x, a, b)\}, \quad \text{for all } x, p \in \mathbf{R}^N,$$

where A, B are compact, \mathcal{T} is closed, and f, g, h are sufficiently smooth but mainly $h(x, a, b) \geq h_0 > 0$. In this framework, we proved a free boundary uniqueness result: There exists at most one pair (U, Ω) where $\Omega \supset \mathcal{T}$ is open and U is continuous in $(\Omega \setminus \mathcal{T}) \cup \partial\mathcal{T}$, is bounded below, and is a viscosity solution of (0.1). In particular, the solution of (0.1) is the value function V of a differential game whose dynamics have vector field f , running cost h , terminal set \mathcal{T} , and final cost g , without any regularity assumption on V .

Our first purpose here is to revisit that result and state it as a free-boundary comparison theorem: We assume that U satisfies the differential equation that appears in (0.1) and the boundary condition on $\partial\mathcal{T}$ as a viscosity subsolution or supersolution, and we state, respectively, that $U \leq V$ or $U \geq V$ in $\Omega \setminus \mathcal{T}$. We also prove the corresponding inequalities between the sets Ω and $\{x: V(x) < +\infty\}$. This result can, in fact, be viewed as a sub- or superoptimality principle, respectively, in the sense of Lions–Souganidis [LSO]. If g is bounded, we also allow more general boundary conditions on $\partial\Omega$, such as $U \equiv c$, where $c > \sup_{\partial\mathcal{T}} g$. Moreover, we prove that the same results also hold if the Hamiltonian can be written as

$$(0.3) \quad \tilde{H}(x, p) := \sup_{a \in A} \inf_{b \in B} \{-f(x, a, b) \cdot p - h(x, a, b)\},$$

i.e., when sup and inf have been interchanged. The proof in this case has some relevant differences with respect to the one in the other case. This is mainly due to the fact that

* Received by the editors April 12, 1991; accepted for publication (in revised form) November 8, 1991.

† Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, 35131 Padova, Italy.

the two representations of the Hamiltonian arise in the context of survival differential games, but (0.2) refers to a pursuit problem in which the minimizing player (or pursuer) is the leader of the game, whereas (0.3) applies to the evasion problem in which the leader is the maximizing player. In fact, it is well known that, in general (if $H \neq \tilde{H}$), there is not a good definition of value function. A lower value V and an upper value \tilde{V} are introduced, and they provide some advantage to the minimizing or, respectively, the maximizing player (evader) (we will give the precise definitions in the next section; for more remarks about the notion of value for differential games, we also refer to [EK1], [Fl], [Fr], and [S2]).

Our comparison theorems have several applications in the theory of controls and differential games. In particular, we can prove rather general relaxation theorems (weak bang-bang principles), give sufficient conditions for the existence of the value, and prove the equivalence among different concepts of value; see also [BS1] and [S2].

We state as a comparison theorem a local uniqueness result in [BS1], which generalizes a previous one from Evans and James [EJ], for the Dirichlet problem

$$(0.4) \quad \begin{aligned} H(x, DU) &= 0 \quad \text{in } B(x_0, r) \setminus \mathcal{T}; \\ U &= g \quad \text{on } \partial\mathcal{T}. \end{aligned}$$

As a consequence of this last result, we then prove that, if (0.4) has continuous viscosity sub- and supersolutions for all $x_0 \in \partial\mathcal{T}$ and some choice of r , possibly depending on x_0 , then (0.1) has a solution pair (Ω, U) , with U continuous in $(\Omega \setminus \mathcal{T}) \cup \partial\mathcal{T}$. This formulation of an existence theorem for a Dirichlet problem associated with a Hamilton–Jacobi equation is new. In fact, in the literature on existence of solutions for such problems, the existence of global viscosity sub- and supersolutions is usually assumed (see [Ba], [CL], [L1], [L2], and [I1]). Finally, for an understanding of the Dirichlet problems for Hamilton–Jacobi equations, we also refer to Soner [Sn], Ishii [I2], and Bardi and Soravia [BS2], and to the references therein.

The second part of this work is devoted to other consequences of the previous uniqueness results. In some papers (see Yong [Y1] and the works cited therein), the notions of local and local asymptotic terminability are introduced for pursuit games, generalizing the usual ones for dynamical systems. These notions are also related to the much-studied problem of stabilization of control systems (see Brockett [Br]). The opposite problems of evadability and strict evadability have also been studied for evasion games (see Yong [Y2] and the references therein). In the present work, we give sufficient conditions for terminability and evadability of systems, which generalize many of the results in [Y1] and [Y2] in two ways. It is well known that the classical sufficient condition for the local asymptotic stability of a dynamical system without any control $y' = f(y)$, where $f(0) = 0$ (here $\mathcal{T} = \{0\}$), prescribes the existence of a Lyapunov function U that satisfies the differential inequality

$$-f(x) \cdot DU(x) > 0$$

in a neighborhood of the origin. In a first result, we show that it is enough to assume that U is continuous and verifies the previous condition in the viscosity sense, and we prove that this result extends to the more general context of differential games, which requires a natural generalization of the classical differential inequality. However, known results for systems without controls (see, e.g., Bhatia, Szego, and Yorke [BSY] and Yorke [Yo]) involve Lyapunov functions that are just semicontinuous by means of certain weak derivatives (for more about weak conditions on U , see also Lakshmikantham and Leela [LL]). These stronger results could, in fact, be obtained in the classical case, also in the context of viscosity solutions, by a result due to Crandall and Lions

[CL]. Some related questions regarding semicontinuous Lyapunov functions and differential inclusions, and therefore control problems, can be found in Aubin [A], Aubin and Cellina [AC], and Aubin and Frankowska [AF].

In a second result, we give explicit conditions on the direction of the vector field f at the points of the boundary of the terminal set. We are able to manage general closed sets, not just closed convex sets in particular subspaces, as in the previous literature. There are two reasons for this. First, we need to construct a Lyapunov function that satisfies the differential inequality in the weak sense of viscosity solutions. Moreover, we employ a generalization of the concept of a normal vector, usually introduced for convex or smooth sets, which seems to be suitable in our framework. We apply a definition of an exterior normal vector to a general closed set \mathcal{T} , which, to our knowledge, was first used by Bony [B]. Roughly, a vector is said to be exterior normal to \mathcal{T} if, in that direction, we can construct a tangent exterior ball to \mathcal{T} . We adopt this definition instead of others that are perhaps more used in the literature because such vectors provide a nice representation of the viscosity sub- and super-differentials of the distance function to \mathcal{T} . More comments about this will follow in Remark 3.8. One of the referees pointed out to us that this definition has been used by Crandall and Newcomb [CN] and Souganidis [So] in the theory of viscosity solutions.

In the last part of § 3 we will develop similar techniques and results for the evasion problem.

1. Preliminaries. In this section, we describe the framework of the problem and the main assumptions that will hold throughout the paper. We consider a dynamical system controlled by two players

$$(1.1) \quad y' = f(y, a, b), \quad y(0) = x \in \mathbf{R}^N,$$

where the vector field $f: \mathbf{R}^N \times A \times B \rightarrow \mathbf{R}^N$ is continuous and satisfies a uniform Lipschitz condition in the state variable

$$|f(x, a, b) - f(z, a, b)| \leq L|x - z| \quad \text{for all } x, z, a, b.$$

We suppose that the control sets A, B are compact subsets of \mathbf{R}^M and that the controls a and b belong, respectively, to the following sets of admissible controls:

$$\mathcal{A} := \{a: \mathbf{R}_+ \rightarrow A \text{ measurable}\}, \quad \mathcal{B} := \{b: \mathbf{R}_+ \rightarrow B \text{ measurable}\}.$$

We denote by $y_x(\cdot; a, b)$, or simply by $y_x(\cdot)$, $y(\cdot)$, a solution of system (1.1) corresponding to a choice of a and b . We are also given a closed target (or terminal set) $\mathcal{T} \in \mathbf{R}^N$ and the two following functions: a running cost $h: \mathbf{R}^N \times A \times B \rightarrow \mathbf{R}$, which we suppose to be continuous and strictly positive, i.e.,

$$(1.2) \quad h(x, a, b) \geq h_0 > 0 \quad \text{for all } x, a, b,$$

and a final cost $g: \mathcal{T} \rightarrow [G; +\infty[$, which will be bounded below ($G > -\infty$) and continuous. For simplicity of notation, we assume that, in (1.2), $h_0 = 1$. This condition on the sign of h can be weakened in some cases, as will be stated in Remark 3.12.

For each choice of a and b , the game starts at the point x at time 0 and ends at the first time the trajectory hits the target, i.e., at the time

$$(1.3) \quad t_x = t_x(a, b) := \inf \{t: y(t) \in \mathcal{T}\} \leq +\infty,$$

where $t_x = +\infty$ if $y_x(\cdot)$ never reaches the target. The payoff of such a trajectory is defined to be

$$(1.4) \quad P(x, a, b) := \int_0^{t_x} h(y_x(t), a(t), b(t)) dt + g(y_x(t_x)).$$

We introduce the set of admissible strategies for the first player by following the ideas of various authors, i.e., Varaiya [V], Roxin [R], and Elliott and Kalton [EK1] (henceforth denoted VREK)

$$\Delta := \{\alpha: \mathcal{B} \rightarrow \mathcal{A}: b(t) = b'(t) \text{ for a.e. } t \leq t'\}$$

$$\text{implies } \alpha[b](t) = \alpha[b'](t) \text{ for a.e. } t \leq t',$$

and we indicate by Γ the corresponding set of strategies for the second player. The lower VREK value is defined by

$$V(x) := \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} P(x, \alpha[b], b),$$

and, in the same way, $\tilde{V}(x) = \sup_{\beta \in \Gamma} \inf_{a \in \mathcal{A}} P(x, a, \beta[a])$ is the upper VREK value. In the particular case of $h \equiv 1$ and $g \equiv 0$, we indicate with T and \tilde{T} the lower and the upper value functions; they are usually called the lower and upper capture time, respectively. The next definitions and results of this section deal with the lower value V , but everything can be translated, with obvious changes, in terms of \tilde{V} .

We also introduce the (lower) capturability set, namely, the set of starting points of the game such that the first player can choose a strategy that forces the system into the target in time less than some positive constant, no matter which control is selected by the second; i.e.,

$$\mathcal{R} := \{x: V(x) < +\infty\}.$$

$\tilde{\mathcal{R}}$ indicates the corresponding set for the upper value. It is not difficult to prove (see [BS1, Lemma 1.1]) that \mathcal{R} is independent of the choice of h and g . More generally, if $C > 0$, we set $\mathcal{R}(C) := \{x: V(x) < C\}$.

We now introduce the following change of variables, or Kruzkov transformation, which is crucial in the remainder of the paper:

$$(1.5) \quad \psi(r) = 1 - \exp(-r).$$

Consider the function defined by $v(x) := \psi(V(x))$ if $x \in \mathcal{R}$ and extended by $v(x) := 1$ if $x \in \mathcal{R}^c$. It is important to note that v is bounded and is itself the (lower) value function of a differential game. In fact, it is easily seen that

$$v(x) := \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} \psi(P(x, \alpha[b], b)) \quad \text{for all } x \in \mathbf{R}^N \setminus \mathcal{T},$$

where the payoff in this case is

$$(1.6) \quad \begin{aligned} \psi(P(x, a, b)) = & \int_0^{t_x} h(y_x(t), a(t), b(t)) \exp\left(-\int_0^t h(y_x(s), a(s), b(s)) ds\right) dt \\ & + \exp\left(-\int_0^{t_x} h(y_x(s), a(s), b(s)) ds\right) \psi(g(y_x(t_x))). \end{aligned}$$

We now briefly recall the definition of a discontinuous viscosity solution of a Hamilton-Jacobi equation as introduced by Ishii [I1], which generalizes the original definition given by Crandall and Lions [CL]. Let $\Omega \in \mathbf{R}^N$ be an open set, and $F: \Omega \times \mathbf{R} \times \mathbf{R}^N \rightarrow \mathbf{R}$ be a continuous function. $u_1, u_2: \Omega \rightarrow \mathbf{R}$ are, respectively, a viscosity subsolution and supersolution of $F(x, u, Du) = 0$ in Ω if u_1 is upper semicontinuous, u_2 is lower semicontinuous, and for all $\varphi \in C^1(\Omega)$ such that $u_1 - \varphi$ attains a local maximum point at y (respectively, $u_2 - \varphi$ attains a local minimum point at y) we have $F(y, u_1(y), D\varphi(y)) \leq 0$ (respectively, $F(y, u_2(y), D\varphi(y)) \geq 0$). The set of these vectors $D\varphi(y)$ is called the superdifferential of u_1 at y and is denoted by $D^+u_1(y)$ (respectively, the subdifferential $D^-u_2(y)$). For an equivalent and more explicit definition of viscosity sub- and superdifferentials, see Crandall, Evans, and Lions [CEL].

A function $u: \Omega \rightarrow \mathbf{R}$ is a viscosity solution of $F(x, u, Du) = 0$ in Ω if the upper and lower semicontinuous envelopes of u

$$u^*(x) := \limsup_{y \rightarrow x} u(y), \quad u_*(x) := \liminf_{y \rightarrow x} u(y)$$

are, respectively, a subsolution and a supersolution.

We denote the relevant Hamiltonians for our problem,

$$(1.7) \quad H(x, p) := \min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot p - h(x, a, b)\},$$

$$(1.8) \quad \mathcal{H}(x, r, p) := \min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot p - h(x, a, b) + (h(x, a, b) - 1)r\},$$

and also define \tilde{H} and $\tilde{\mathcal{H}}$ by interchanging min and max in (1.7) and (1.8).

DEFINITION 1.1. A function $\sigma: \mathcal{A} \times \mathcal{B} \rightarrow \mathbf{R}_+$ is nonanticipating if $a(t) = a'(t)$ and $b(t) = b'(t)$ for almost every $0 \leq t \leq \sigma(a, b)$ imply that $\sigma(a, b) = \sigma(a', b')$.

The following general formulation of the dynamic-programming principle, which holds for differential games, is well known (see Evans and Ishii [EI] and Elliott and Kalton [EK2]).

PROPOSITION 1.2. Let $x \in \mathbf{R}^N \setminus \mathcal{T}$ and σ be a nonanticipating function such that $\sigma(a, b) \leq t_x(a, b)$ for all a, b . Then

$$V(x) := \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} \left\{ \int_0^\sigma h(y_x(t), \alpha[b](t), b(t)) dt + V(y_x(\sigma(\alpha[b], b))) \right\}.$$

The next result is based on the dynamic-programming principle and relates the value functions V , v , and the Hamiltonians in (1.7) and (1.8). It is proved by combining the arguments in [ES] and [I2]. The corresponding statement for the upper value \tilde{V} uses the Hamiltonians \tilde{H} and $\tilde{\mathcal{H}}$.

PROPOSITION 1.3. (i) If \mathcal{R} is open and V is locally bounded, then V is a viscosity solution of

$$(1.9) \quad H(x, DV) = 0 \quad \text{in } \mathcal{R} \setminus \mathcal{T};$$

(ii) v is a viscosity solution of

$$(1.10) \quad v(x) + \mathcal{H}(x, v(x), Dv(x)) = 0 \quad \text{in } \mathbf{R}^N \setminus \mathcal{T}.$$

We recall that (1.9) and (1.10) are usually called the Isaacs equations of the lower games with payoff given in (1.4) and (1.6), respectively, and that the upper value functions satisfy the above equations with \tilde{H} and $\tilde{\mathcal{H}}$ in place of H and \mathcal{H} .

We now introduce another notation: Let $\Omega \in \mathbf{R}^N$ be an open set and let $x \in \Omega$. For any nonanticipating function $\sigma(a, b) \leq \tau_x(a, b)$, where $\tau_x(a, b)$ denotes the first exit time from Ω , i.e., $\tau_x(a, b) = \inf \{t \geq 0 : y_x(t) \in \Omega^c\}$, and for all $a \in \mathcal{A}$, $b \in \mathcal{B}$, $u: \partial\Omega \rightarrow \mathbf{R}$, we denote

$$(1.11) \quad \begin{aligned} Q(\sigma, x, a, b, u) &:= \int_0^\sigma h(y_x(t), a(t), b(t)) \exp \left(- \int_0^t h(y_x(s), a(s), b(s)) ds \right) dt \\ &+ \exp \left(- \int_0^\sigma h(y_x(s), a(s), b(s)) ds \right) u(y_x(\sigma)) \\ &= 1 - \exp \left(- \int_0^\sigma h(y_x(t), a(t), b(t)) dt \right) (1 - u(y_x(\sigma))). \end{aligned}$$

The next result is a modification of Proposition 2.3 in [BS1] and generalizes similar statements in Evans and Ishii [EI] and in Lions and Souganidis [LSO].

PROPOSITION 1.4. Let $\Omega \subset \mathbf{R}^N$ be an open set, and let $\tau_x(a, b)$ denote the first exit time from Ω . If $u \in C(\Omega)$ is a bounded viscosity subsolution (respectively, supersolution) of

$$u(x) + \mathcal{H}(x, u(x), Du(x)) = 0 \quad \text{in } \Omega$$

then $u(x) \leq$ (respectively, \geq) $\inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} Q(\tau_x, x, \alpha[b], b, u)$. (Note that only values of u on $\partial\Omega$ appear in the right-hand side of the inequality.)

Proof. Assume that u is a subsolution. Given $\varepsilon > 0$, we consider the open set

$$\Omega_\varepsilon^\wedge := \{x \in \Omega: \text{dist}(x, \partial\Omega) > \varepsilon, |x| < 1/\varepsilon\}.$$

It is well known (largely through Sard's lemma) that, for each $\varepsilon > 0$, there exists, as $\varepsilon \rightarrow 0$, an increasing family of sets $\Omega_\varepsilon, \Omega_{\varepsilon/2}^\wedge \in \Omega_\varepsilon^\wedge$ such that $\partial\Omega_\varepsilon$ is a smooth manifold. Now we define the function

$$u^\varepsilon(x) := \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} Q(\tau_x^\varepsilon, x, \alpha[b], b, u),$$

where τ_x^ε is the first exit time from Ω_ε . By Ishii [I2] (see also Proposition 1.3), it follows that u^ε satisfies in the viscosity sense

$$v(x) + \mathcal{H}(x, v(x), Dv(x)) = 0 \quad \text{in } \Omega_\varepsilon,$$

$$v = u \quad \text{or} \quad v(x) + \mathcal{H}(x, v(x), Dv(x)) = 0 \quad \text{on } \partial\Omega_\varepsilon.$$

By the comparison Theorem 1.1 in [BS3], we can deduce that

$$u \leq (u^\varepsilon)_* \leq u^\varepsilon \quad \text{in } \Omega_\varepsilon.$$

Now the proof follows the arguments of Theorem 4.1 in [E1], to pass to the limit as $\varepsilon \rightarrow 0$ and reach the conclusion. \square

We introduce more notation: Let $r > 0$; then $B(x, r) := \{y \in \mathbf{R}^N: |x - y| < r\}$, and for $x \in \mathbf{R}^N$, $B(X, r) := \{x: \text{dist}(x, X) < r\}$.

DEFINITION 1.5. Let $\mathcal{T} \subset \mathbf{R}^N$ be a closed set. A vector $\nu \in \mathbf{R}^N$, $|\nu| = 1$ is said to be exterior normal to \mathcal{T} at $x \in \partial\mathcal{T}$ if there exists $\varepsilon > 0$ such that

$$B(x + \varepsilon\nu, \varepsilon) \cap \mathcal{T} = \emptyset.$$

We also denote $d(x) := \text{dist}(x, \mathcal{T})$, the distance function from the target. It is possible to compute explicitly the viscosity sub- and superdifferentials of the function d . This result is contained in the next proposition and will also motivate Definition 1.5.

PROPOSITION 1.6. Let $x \in \mathbf{R}^N \setminus \mathcal{T}$.

(i) Assume that there exists a unique $z \in \partial\mathcal{T}$ such that $d(x) = |x - z|$, then $D^-d(x) = \{(x - z)/|x - z|\}$. Otherwise $D^-d(x)$ is empty;

(ii) $D^+d(x) = \text{co} \{(x - z)/|x - z|: z \in \partial\mathcal{T}, d(x) = |x - z|\}$.

In particular, the set $\{z \in \partial\mathcal{T}: |x - z| = d(x)\}$ consists of just one point \bar{z} if and only if d is differentiable at x . In this case, we have $Dd(x) = (x - \bar{z})/|x - \bar{z}|$.

Proof. Statement (i) substantially follows from the fact that d is a viscosity solution of $|Dd(x)| = 1$ in $\mathbf{R}^N \setminus \mathcal{T}$. Statement (ii) is a consequence of Proposition 1 in Souganidis [So], Theorem 3I in Rockafellar [Ro], and Theorem 2.1 in Clarke [C]. \square

In the following, C will indicate a generic positive constant.

2. Comparison results for the free-boundary problem. This section concerns several comparison results, in local or global form, between the value functions of the differential game introduced in § 1 and continuous viscosity sub- or supersolutions of the corresponding Isaacs equations, assuming that they satisfy suitable boundary conditions. As a consequence of one of these results, we prove the existence of a continuous

viscosity solution to the Dirichlet problem (0.1), assuming the existence of only local sub- and supersolutions.

DEFINITION 2.0. Given an open set $\Omega \supset \mathcal{T}$ and a constant $U_0 \in \mathbf{R} \cup \{+\infty\}$, we say that a function $U \in C((\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T})$ satisfies the boundary condition (BC) if $U(x) < U_0$ in $(\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T}$ and

$$U(x) \rightarrow U_0 \quad \text{as } x \rightarrow x_0 \in \partial \Omega.$$

The first result concerns the upper value function.

THEOREM 2.1. Let $\Omega \supset \mathcal{T}$ be an open set and $U \in C((\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T})$ be bounded below, satisfying (BC) and

$$\tilde{H}(x, DU) \geq 0 \quad \text{in } \Omega \setminus \mathcal{T},$$

$$U \geq g \quad \text{on } \partial \mathcal{T}$$

in the viscosity sense. Then $U \geq \tilde{V}$ in $(\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T}$ and $\Omega \subset \tilde{\mathcal{R}}(U_0)$.

Proof. Let $u(x) := \psi(U(x))$; then we can extend its definition so that $u \in C(\overline{\Omega \setminus \mathcal{T}})$, is bounded and satisfies in the viscosity sense the following inequality:

$$u(x) + \tilde{\mathcal{H}}(x, u(x), Du(x)) \geq 0 \quad \text{in } \Omega \setminus \mathcal{T}.$$

Therefore, by Proposition 1.4 applied to the upper Isaacs equation, it follows that

$$(2.0) \quad u(x) \geq \sup_{\beta \in \mathbf{I}^+} \inf_{a \in \mathcal{A}} Q(\tau_x, x, \alpha[b], b, u),$$

where τ_x is the first exit time from $\Omega \setminus \mathcal{T}$. Let $x \in \Omega$, then $u(x) < \psi(U_0)$, and we can choose $\varepsilon > 0$ satisfying the inequality $2\varepsilon < \psi(U_0) - u(x)$. Thus, by (2.0), for all β , we can select a_β such that

$$(2.1) \quad 1 > \psi(U_0) - \varepsilon \geq u(x) + \varepsilon \geq Q(\tau_x, x, a_\beta, \beta[a_\beta], u).$$

Inequality (2.1) implies that $\tau_x = t_x$ because, if $\tau_x(a, b) = +\infty$, then the right-hand side of (2.1) is equal to unity. On the other hand, if $y_x(\tau_x) \in \partial \Omega$ then we easily obtain

$$\psi(U_0) - \varepsilon \geq \psi\left(\int_0^{\tau_x} h \, dt + U_0\right) \geq \psi(\tau_x + U_0),$$

which is a contradiction. Thus, since $u(z) \geq \psi(g(z))$ at $z \in \partial \mathcal{T}$, by (2.1) we get $u(x) + \varepsilon \geq \psi(P(x, a_\beta, \beta[a_\beta]))$ and then $u(x) \geq \tilde{v}(x)$, which implies both the conclusions. \square

The proof of the counterpart of Theorem 2.1 for subsolutions is not so easy. We need first to prove the following lemma, which roughly states that for a fixed $x \in \mathbf{R}^N \setminus \mathcal{T}$, if we are given two nonanticipating functions such that $\sigma(a, b) \leq \tau(a, b)$, a strategy β_1 , and a control a_1 that are almost optimal for the value function \tilde{v} up to time σ , then we can modify the definitions of β_1 and a_1 in the interval $[\sigma, +\infty[$ so that they become almost optimal up to time τ .

LEMMA 2.2. Let $x \in \mathbf{R}^N \setminus \mathcal{T}$ and let σ_x, τ_x be nonanticipating functions such that $\sigma_x(a, b) \leq \tau_x(a, b)$ for all a, b . Assume that for all β_1 there exists a_1 , which verifies

$$(2.2) \quad Q(\sigma_x, x, a_1, \beta_1[a_1], \tilde{v}) < C.$$

If, for a given β_2 we define

$$\beta[a](t) = \begin{cases} \beta_1[a](t) & \text{if } t \in [0, \sigma(a, \beta_1[a])], \\ \beta_2[a(\cdot - \sigma)](t - \sigma) & \text{elsewhere,} \end{cases}$$

then there exists $a_\beta, a_\beta = a_1$ for $t \leq \sigma(a_1, \beta_1[a_1])$ such that $Q(\tau_x, x, a_\beta, \beta[a_\beta], \tilde{v}) < C$.

Proof. We apply the dynamic programming principle to $\tilde{v}(y_x(\sigma_x))$ in (2.2) with respect to the nonanticipating function $\rho(a, b) := \tau_x(a', b') - \sigma_x(a_1, b_1)$, where $b_1 = \beta_1[a_1]$ and $a'(t) = a_1(t)$, $b'(t) = b_1(t)$ if $t \leq \sigma_x(a_1, b_1)$, $a'(t) = a(t-s)$, $b'(t) = b(t-s)$ elsewhere. Then, for all β_2 , there exists a_2 such that (we drop the x in σ_x)

$$\int_0^\sigma h \exp\left(-\int_0^t h ds\right) dt + \exp\left(-\int_0^\sigma h dt\right) Q(\rho, y(\sigma), a_1, \beta_1[a_1], \tilde{v}) < C.$$

It is now sufficient to define

$$a_\beta(t) = \begin{cases} a_1(t) & \text{if } t \in [0, \sigma(a_1, \beta_1[a_1])], \\ a_2(t-\sigma) & \text{elsewhere} \end{cases}$$

and to apply an easy change of variables to obtain the result. \square

We now want to prove the comparison theorem.

THEOREM 2.3. *Let $\Omega \supset \mathcal{T}$ be an open set and $U \in C((\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T})$ be bounded below, satisfying (BC) and*

$$\begin{aligned} \tilde{H}(x, DU) &\leq 0 \quad \text{in } \Omega \setminus \mathcal{T}, \\ U &\leq g \quad \text{on } \partial \mathcal{T} \end{aligned}$$

in the viscosity sense. Then $U \leq \tilde{V}$ in $(\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T}$ and $\Omega \supset \tilde{\mathcal{R}}(U_0)$.

Proof. First step. We prove that $\tilde{\mathcal{R}}(U_0) \cap \partial \Omega = \emptyset$. To this end, we assume by contradiction that there exists $x \in \tilde{\mathcal{R}}(U_0) \cap \partial \Omega$. Fix $\varepsilon > 0$ such that $\tilde{v}(x) < \psi(U_0) - \varepsilon$, and observe that, if $\psi(P(x, a, b)) < \psi(U_0) - \varepsilon$, then by (1.2) and $g \geq G$, we can choose a constant $C_1 > 0$, independent of the choice of a and b such that $t_x(a, b) \leq C_1$, which implies by Gronwall's lemma that we can find $R > 0$ such that

$$|y_x(t)| \leq R \quad \text{for all } t \leq t_x.$$

Let $C_2 := \sup_{B(0, R) \times A \times B} |f(z, a, b)|$, $\|\psi(g)\|_R := \sup\{|\psi(g(x))|; x \in \partial \mathcal{T} \cap B(0, R)\}$, $0 < \eta < \min\{\varepsilon, \psi(U_0) - \|\psi(g)\|_R\}$. We now choose $\delta > 0$ such that $\text{dist}(z, \partial \Omega) \leq \delta$ and $z \in \Omega \cap B(0, R)$ imply

$$(2.3) \quad u(z) := \psi(U(z)) > \psi(U_0) - \eta > \|\psi(g)\|_R$$

and then $z \notin \partial \mathcal{T}$.

We now define $\sigma_x(a, b)$ as the first exit time of $y_x(\cdot; a, b)$ from $\mathbf{R}^N \setminus \Omega_\delta$, where $\Omega_\delta := \{x \in \Omega; \text{dist}(x, \partial \Omega) \geq \delta\}$. By the dynamic programming principle (Proposition 1.2), we get

$$\tilde{v}(x) = \sup_{\beta \in \Gamma} \inf_{a \in \mathcal{A}} \{Q(\sigma_x, x, a, \beta[a], \tilde{v})\},$$

and then, by the choice of ε , for all β_1 , we can find a_1 such that

$$(2.4) \quad Q(\sigma_x, x, a_1, \beta_1[a_1], \tilde{v}) < \psi(U_0) - \varepsilon$$

and, as a consequence,

$$\begin{aligned} \tilde{v}(y(\sigma)) &\leq 1 - \exp\left(-\int_0^\sigma h ds\right) (1 - \tilde{v}(y(\sigma))) \\ (2.5) \quad &= Q(\sigma_x, x, a_1, \beta_1[a_1], \tilde{v}) < \psi(U_0) - \varepsilon. \end{aligned}$$

We fix β_1 and observe that $z := y(\sigma) \in \partial\Omega_\delta$ and $\text{dist}(z, \partial\Omega) = \delta$. We again apply the dynamic programming principle to $\tilde{v}(z)$ with the nonanticipating function $\tau_z(a, b)$, indicating the first exit time from $\Omega \setminus \mathcal{T}$, and get

$$\hat{v}(z) = \sup_{\beta \in \Gamma} \inf_{a \in \mathcal{A}} \{Q(\tau_z, z, a, \beta[a], \tilde{v})\}.$$

By the previous formula, Lemma 2.2, and (2.4), we deduce that for all β_2 there exists a_2 such that, if we construct β and a_β as in the lemma (in which $\sigma_x + \tau_z$ plays the part of τ), then

$$(2.6) \quad Q(\sigma_x + \tau_z, x, a_\beta, \beta[a_\beta], \tilde{v}) < \psi(U_0) - \varepsilon$$

and, thus, since the running cost is positive,

$$(2.7) \quad Q(\tau_z, z, a_2, \beta_2[a_2], \tilde{v}) < \psi(U_0) - \varepsilon,$$

$$(2.8) \quad \psi(P(x, a_{\beta'}, \beta'[a_{\beta'}], u)) < \psi(U_0) - \varepsilon$$

for a suitable strategy β' and control $a_{\beta'}$ constructed modifying β, a_β in $[\sigma_x + \tau_x, +\infty[$, as in Lemma 2.2, where, in this case, t_x plays the part of τ and $\sigma_x + \tau_x$ plays the part of σ .

Observe now that the function u defined in (2.3) is bounded, can be extended to a continuous function in $\overline{\Omega \setminus \mathcal{T}}$, and satisfies in the viscosity sense

$$u(x') + \tilde{\mathcal{H}}(x', u(x'), Du(x')) \leq 0 \quad \text{in } \Omega \setminus \mathcal{T}.$$

Therefore, by Proposition 1.4 we have that

$$u(x') \leq \sup_{\beta \in \Gamma} \inf_{a \in \mathcal{A}} \{Q(\tau_{x'}, x', a, \beta[a], u)\} \quad \text{in } \Omega \setminus \mathcal{T},$$

where $\tau_{x'}(a, b)$ is the first exit time from $\Omega \setminus \mathcal{T}$. Now (2.3) and the previous formula imply that there exists $\tilde{\beta}$ such that

$$Q(\tau_z, z, a, \tilde{\beta}[a], u) > \psi(U_0) - \eta \quad \text{for all } a \in \mathcal{A}.$$

If $t_z(a, \tilde{\beta}[a]) = \tau_z(a, \tilde{\beta}[a])$ for some a , since $u \leq \psi(g)$ on $\partial\mathcal{T}$, we obtain

$$(2.9) \quad \psi(P(z, a, \tilde{\beta}[a])) > \psi(U_0) - \eta.$$

If instead $t_z > \tau_z$, then

$$(2.10) \quad \begin{aligned} Q(\tau_z, z, a, \tilde{\beta}[a], u) &= 1 - \exp\left(-\int_0^{\tau_z} h \, ds\right) (1 - u(y(\tau_z))) \\ &= \psi\left(\int_0^{\tau_z} h \, dt + U(y_x(\tau_z))\right) \geq \psi(\tau_z + U_0). \end{aligned}$$

Now we specialize $\beta_2 = \tilde{\beta}$, $a = a_2$, and observe that (2.7) contradicts (2.9). Therefore, we obtain $\tau_z < t_z$. By (2.7) and (2.10), we now obtain

$$(2.11) \quad \begin{aligned} \psi(U_0) - \tilde{v}(y(\tau_z)) &= u(y(\tau_z)) - \tilde{v}(y(\tau_z)) \\ &\geq \exp\left(-\int_0^{\tau_z} h \, ds\right) \left(u(y(\tau_z)) - \tilde{v}(y(\tau_z))\right) \\ &= Q(\tau_z, z, a_2, \tilde{\beta}[a_2], u) - Q(\tau_z, z, a_2, \tilde{\beta}[a_2], \tilde{v}) \\ &\geq \psi(\tau_z + U_0) - \psi(U_0) + \varepsilon. \end{aligned}$$

By (2.8) we also have $\delta \leq |y(\tau_z) - z| \leq \int_0^{\tau_z} |f| ds \leq C_2 \tau_z$ and then $\tau_z \geq \delta / C_2$, which, in view of (2.11), implies

$$\tilde{v}(y(\tau_z)) \leq \psi(U_0) - \varepsilon - (\psi(\delta / C_2 + U_0) - \psi(U_0)).$$

Therefore the point $x_1 := y_z(\tau_z) = y_x(\sigma_x + \tau_z) \in \partial\Omega \cap \tilde{\mathcal{R}}(U_0)$, lies in $B(0, R)$ by (2.8), and $\tilde{v}(x_1) < \psi(U_0) - \varepsilon - \rho$, where the positive constant ρ depends on δ .

We now repeat the procedure above, starting from $x_1 = y_x(\sigma_x + \tau_z)$, and use Lemma 2.2 and the fact that the trajectory leading from x to x_1 can be extended to a trajectory leading to the target whose payoff is $\leq \psi(U_0) - \varepsilon$ (i.e., $t_x \leq C_1$) by (2.6) and (2.10). Then we can find another point $x_2 \in \partial\Omega \cap \tilde{\mathcal{R}}(U_0)$ on a trajectory performing the same property, which is therefore completely contained in the ball $B(0, R)$. Thus, the value δ does not change, the decrement ρ is constant, and $\tilde{v}(x_2) < \psi(U_0) - \varepsilon - 2\rho$. We proceed in this way, but since \tilde{v} is bounded, after a finite number of steps we find the required contradiction.

Second step. Let $x \in \tilde{\mathcal{R}}(U_0)$; we apply the dynamic-programming principle and get

$$\tilde{v}(x) = \sup_{\beta \in \Gamma} \inf_{a \in \mathcal{A}} \{Q(\tau_x, x, a, \beta[a], \tilde{v})\},$$

where τ_x is the first exit time from $\Omega \setminus \mathcal{T}$ if $x \in \Omega \setminus \mathcal{T}$ or the exit time from $\mathbf{R}^N \setminus \bar{\Omega}$ if $x \in \mathbf{R}^N \setminus \bar{\Omega}$.

We choose $\varepsilon > 0$ such that $\tilde{v}(x) + \varepsilon < \psi(U_0) \leq 1$; then, for all β , there exists a_β such that

$$Q(\tau_x, x, a_\beta, \beta[a_\beta], \tilde{v}) < \tilde{v}(x) + \varepsilon,$$

and therefore we conclude that $\tau_x \neq +\infty$ and then, since $\tilde{v}(y_x(\tau_x)) \leq Q(\tau_x, x, a_\beta, \beta[a_\beta], \tilde{v})$, that $y(\tau_x) \in \tilde{\mathcal{R}}(U_0)$ which implies $y(\tau_x) \notin \partial\Omega$. This gives a contradiction for $x \in \mathbf{R}^N \setminus \bar{\Omega}$ and then $\tilde{\mathcal{R}}(U_0) \subset \Omega$. For $x \in \Omega$, we instead obtain $\tau_x = t_x$. Therefore, since $u \leq \psi(g)$ on $\partial\mathcal{T}$, we get $Q(\tau_x, x, a_\beta, \beta[a_\beta], u) < \tilde{v}(x) + \varepsilon$, which implies, by Proposition 1.4, that $u \leq \tilde{v}$ in $\tilde{\mathcal{R}}(U_0)$. Now it is obvious by the definitions that the same inequality holds in $\Omega \setminus \mathcal{T}$. \square

The next uniqueness result is a consequence of both Theorems 2.1 and 2.3.

COROLLARY 2.4 (uniqueness for the free-boundary problem). *Assume that there exists one pair (Ω, U) such that $\Omega \supset \mathcal{T}$ is open, $U \in C((\Omega \setminus \mathcal{T}) \cup \partial\mathcal{T})$ is bounded below, satisfying*

$$\tilde{H}(x, DU) = 0 \quad \text{in } \Omega \setminus \mathcal{T},$$

$$U = g \quad \text{on } \partial\mathcal{T}$$

in the viscosity sense and (BC). Then $\Omega = \tilde{\mathcal{R}}(U_0)$ and $U = \tilde{V}$ in $(\Omega \setminus \mathcal{T}) \cup \partial\mathcal{T}$.

We now complete the framework by stating the corresponding propositions for the lower value.

THEOREM 2.5. *Let $\Omega \supset \mathcal{T}$ be an open set and $U \in C((\Omega \setminus \mathcal{T}) \cup \partial\mathcal{T})$ be bounded below, satisfying*

$$H(x, DU) \geq 0 \quad \text{in } \Omega \setminus \mathcal{T},$$

$$U \geq g \quad \text{on } \partial\mathcal{T}$$

in the viscosity sense and (BC). Then $U \geq V$ in $(\Omega \setminus \mathcal{T}) \cup \partial\mathcal{T}$ and $\Omega \subset \mathcal{R}(U_0)$.

The proof of the previous result follows that of Theorem 2.1. The proof of the next theorem instead presents some relevant differences from the proof of Theorem 2.3 and is the same as the proof of Theorem 3.1 in [BS1].

THEOREM 2.6. *Let $\Omega \supset \mathcal{T}$ be an open set and $U \in C((\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T})$ be bounded below, satisfying*

$$H(x, DU) \leq 0 \quad \text{in } \Omega \setminus \mathcal{T},$$

$$U \leq g \quad \text{on } \partial \mathcal{T}$$

in the viscosity sense and (BC). Then $U \leq V$ in $(\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T}$ and $\Omega \supset \mathcal{R}(U_0)$.

COROLLARY 2.7. *Assume that there exists one pair (Ω, U) such that $\Omega \supset \mathcal{T}$ is open, $U \in C((\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T})$ is bounded below, satisfying*

$$H(x, DU) = 0 \quad \text{in } \Omega \setminus \mathcal{T},$$

$$U = g \quad \text{on } \partial \mathcal{T}$$

in the viscosity sense and (BC). Then $\Omega = \mathcal{R}(U_0)$ and $U = V$ in $(\Omega \setminus \mathcal{T}) \cup \partial \mathcal{T}$.

COROLLARY 2.8 (comparison of value functions and existence of value). *Assume that either V or \tilde{V} is continuous at the points of $\partial \mathcal{T}$. Then $\tilde{\mathcal{R}} \subset \mathcal{R}$ and $V(x) \leq \tilde{V}(x)$ for all $x \in \tilde{\mathcal{R}}$. If, moreover, $H = \tilde{H}$, then $\tilde{\mathcal{R}} = \mathcal{R}$ and $V = \tilde{V}$.*

Proof. We first observe that $H \geq \tilde{H}$. Moreover, by Theorem 6.1 in [BS1], if V is continuous at the points of $\partial \mathcal{T}$, then it is continuous in \mathcal{R} , \mathcal{R} is open, and the pair (\mathcal{R}, V) satisfies the boundary-value problem (0.1). The corresponding result holds for the upper value \tilde{V} . Therefore, it is sufficient to apply Theorem 2.3 or Theorem 2.5, respectively, if V or \tilde{V} is continuous at the points of $\partial \mathcal{T}$. If, moreover, $H = \tilde{H}$, we can apply Corollary 2.4 or Corollary 2.7, respectively. \square

We now turn to local uniqueness results. We give here the following statement, whose proof is the same as the proof of Theorem 4.1 in [BS1], but uses Proposition 1.4 instead of Proposition 2.3 in [BS2] and also holds for the upper Isaacs equation with obvious changes.

THEOREM 2.9. *Let $x_0 \in \partial \mathcal{T}$, $r > 0$, and $g \equiv 0$. If $U \in C(\overline{B(x_0, r) \setminus \mathcal{T}})$ is a nonnegative viscosity subsolution (respectively, supersolution) of*

$$H(x, DU) = 0 \quad \text{in } B(x_0, r) \setminus \mathcal{T}$$

and $U = 0$ on $\partial \mathcal{T} \cap B(x_0, r)$, then there exists $r' \leq r$ such that $U(x) \leq V(x)$ (respectively, $U(x) \geq V(x)$) for all $x \in B(x_0, r') \setminus \mathcal{T}$.

Remark 2.10. The previous result still holds if we have a general final cost g on $\partial \mathcal{T}$. In this case, we need to assume, instead of requiring U to be nonnegative, the following conditions:

$$(2.12) \quad U(x_0) < r/M_r + \min_{\partial(B(x_0, r) \setminus \mathcal{T})} U,$$

where $M_r := \max \{|f(x, a, b)| : x \in B(x_0, r), a \in A, b \in B\}$, and, if U is subsolution,

$$(2.13) \quad g(x) \geq \min_{\partial(B(x_0, r) \setminus \mathcal{T})} U \quad \text{for all } x \in \partial \mathcal{T}.$$

A different formulation of local comparison theorems can be found in Theorem 3.5. It is now easy to prove the following existence result for problem (0.1), which we state for only the lower Isaacs equation.

THEOREM 2.11. *Assume that for all $x_0 \in \partial\mathcal{T}$ there exist $r > 0$ and two functions $U_1, U_2 \in C(\overline{B(x_0, r)} \setminus \mathcal{T})$, respectively, viscosity super- and subsolutions of the Dirichlet problem*

$$\begin{aligned} H(x, DU) &= 0 \quad \text{in } B(x_0, r) \setminus \mathcal{T}, \\ U &= g \quad \text{on } \partial\mathcal{T} \cap B(x_0, r). \end{aligned}$$

If, moreover, U_1, U_2 satisfy (2.12), U_2 also verifies (2.13) and $U_1(x_0) = U_2(x_0) = g(x_0)$, then V is continuous in $(\mathcal{R} \setminus \mathcal{T}) \cup \partial\mathcal{T}$, and the pair (\mathcal{R}, V) solves the free-boundary problem (0.1).

Proof. By Theorem 2.9 and Remark 2.10, it follows that, for all $x_0 \in \partial\mathcal{T}$ and for some $r'(x_0) > 0$,

$$U_2(x) \leq V(x) \leq U_1(x) \quad \text{in } B(x_0, r') \setminus \mathcal{T}.$$

Then V is continuous at $x_0 \in \partial\mathcal{T}$. By Theorem 6.1 in [BS1], this is enough to conclude the proof. \square

3. Pursuit and evasion problems. In this section, we study the problems of terminability and evadability for a dynamical system with two competitive controls. We first introduce the following definitions.

DEFINITION 3.0. The differential game is stable if for all neighborhoods \mathcal{U} of \mathcal{T} we can find a neighborhood \mathcal{V} of \mathcal{T} which verifies the following condition: For all $x \in \mathcal{V}$, there is $\alpha \in \Delta$ such that $y_x(t) \in \mathcal{U}$ for all $b \in \mathcal{B}$ and $t \in [0, t_x]$.

DEFINITION 3.1. The differential game is locally asymptotically terminable if there exists an open set $\Omega \supset \mathcal{T}$ such that, for all $x \in \Omega \setminus \mathcal{T}$, there is $\alpha \in \Delta$ that verifies $t_x(\alpha[b], b) < +\infty$ or $\lim_{t \rightarrow +\infty} d(y_x(t)) = 0$ for all $b \in \mathcal{B}$.

DEFINITION 3.2. The differential game is locally terminable if there exists an open set $\Omega \supset \mathcal{T}$ such that $\Omega \subset \mathcal{R}$, i.e., for all $x \in \Omega \setminus \mathcal{T}$, there are $C > 0$ and $\alpha \in \Delta$ that satisfy $t_x(\alpha[b], b) \leq C$ for all $b \in \mathcal{B}$.

The previous notions generalize the usual ones for dynamical systems and have been used in control theory and in differential games by previous authors (see, e.g., Lee and Markus [LM], Yong [Y1], and the references therein). We note that, in the case of control systems, the asymptotical terminability (controllability) is, in general, only a necessary condition for the stabilization of the system, whose definition requires the existence of a feedback control. For the relationships between the two approaches, we refer to Brockett [Br]. If, in the definitions above, the set Ω is all of \mathbf{R}^N , we say that the game is globally asymptotically terminable or, respectively, globally terminable. This problem is obviously related to the properties of the function T and its Hamiltonian H .

We also want to study the counterpart of the terminability problem, that is, the evadability of a differential game from a closed set \mathcal{T} . This is a much-studied topic in the literature (see Yong [Y2] and the references therein). We introduce the following two definitions.

DEFINITION 3.3. The game is evadable if, for all $x \in \mathbf{R}^N \setminus \mathcal{T}$, there exists $\beta \in \Gamma$ such that $d(y_x(t; a, \beta[a])) > 0$ for all $a \in \mathcal{A}$ and $t \geq 0$.

DEFINITION 3.4. The game is strictly evadable if, for all $x \in \mathbf{R}^N \setminus \mathcal{T}$, there exist $\delta > 0$ and $\beta \in \Gamma$ such that $d(y_x(t; a, \beta[a])) \geq \delta$ for all $a \in \mathcal{A}$ and $t \geq 0$.

This problem is related to the properties of the function \tilde{T} and its Hamiltonian \tilde{H} . For example, it is clear that a necessary, but not sufficient, condition for the evadability is $\tilde{\mathcal{R}} = \mathcal{T}$ or, similarly, $\tilde{T}(x) = +\infty$ for all $x \in \mathbf{R}^N \setminus \mathcal{T}$. To escape from this apparent problem in the study of the evasion game by dynamic-programming methods, we need the change of variables in (1.5).

3.1. Pursuit problem. We start providing sufficient conditions for terminability by assuming the existence of a suitable Lyapunov function. This function will be assumed continuous, and the proofs will use the comparison theorems for viscosity solutions of the Isaacs equations proved in § 2.

THEOREM 3.5. *Let $\Omega \supset \mathcal{T}$ be an open set. Assume that there exist $\delta > 0$ and $U \in C(\overline{\Omega \setminus \mathcal{T}})$, bounded below, that satisfy in the viscosity sense*

$$\min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot DU(x)\} \geq \delta \quad \text{in } \Omega \setminus \mathcal{T},$$

$$U = 0 \quad \text{on } \partial \mathcal{T},$$

and such that $\inf_{\partial \Omega} U > 0$. Consider the open set $\Omega' := \{x: U(x) < \inf_{\partial \Omega} U\}$; then the game is terminable in Ω' , and, moreover, we have the estimate $T(x) \leq U(x)/\delta$ for all $x \in \Omega' \setminus \mathcal{T}$. Thus the game is stable.

Proof. Define $W(x) := U(x)/\delta$; then W satisfies in the viscosity sense the following problem:

$$H(x, DW(x)) \geq 0 \quad \text{in } \Omega' \setminus \mathcal{T},$$

$$W = 0 \quad \text{on } \partial \mathcal{T},$$

$$W(x) = \inf_{\partial \Omega} W \quad \text{on } \partial \Omega',$$

where H is defined in (1.7) with $h \equiv 1$ and $W(x) < \inf_{\partial \Omega} W$ for $x \in \Omega'$. If we now apply Theorem 2.5, we get the local terminability. The estimate $T(x) \leq U(x)/\delta$ implies that T is continuous on $\partial \mathcal{T}$ if $U = 0$ on $\partial \mathcal{T}$ and immediately gives the stability of the game. In fact, given an open neighborhood \mathcal{U} of \mathcal{T} , let $x_0 \in \partial \mathcal{T}$, $r > 0$ such that $B(x_0, r) \in \mathcal{U}$ and $M := \max_{B(x_0, r) \times A \times B} |f(x, a, b)|$. We select $0 < \varepsilon \leq r/2M$ and r' such that $r' \leq r/2$ and $T(x) < \varepsilon$ for all $x \in B(x_0, r')$. By definition, given $x \in B(x_0, r')$, there is $\alpha \in \Delta$ such that, for all $b \in \mathcal{B}$, we have $t_x < \varepsilon$ and then

$$|y_x(t) - x_0| \leq tM \leq r/2 \quad \text{for all } t \in (0, t_x);$$

that is, $y_x(t) \in B(x_0, r)$ for $t \in (0, t_x)$. \square

In view of Theorem 2.9, by using a similar proof, we can also give a local version of the previous theorem.

THEOREM 3.6. *Let $x_0 \in \partial \mathcal{T}$, $r > 0$ and $U \in C(\overline{B(x_0, r) \setminus \mathcal{T}})$ be a nonnegative function that satisfies in the viscosity sense*

$$\min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot DU(x)\} \geq \delta \quad \text{in } B(x_0, r) \setminus \mathcal{T},$$

$$U = 0 \quad \text{on } \partial \mathcal{T};$$

then there exists $r' \leq r$ such that $T(x) \leq U(x)/\delta$ for all $x \in B(x_0, r') \setminus \mathcal{T}$. Therefore, the game is stable and terminable in $B(x_0, r') \setminus \mathcal{T}$.

We now want to specialize the above results by giving explicit conditions on the vector field f at the points of the boundary of the target \mathcal{T} , which, we recall, is a general closed set, and we want to prove that, under such conditions, the distance function $d(\cdot)$ is the required Lyapunov function.

COROLLARY 3.7. *Let $x_0 \in \partial \mathcal{T}$. Assume that the following condition holds: There exist $\delta, r > 0$ such that*

$$(3.1) \quad F(x, \nu) := \max_{b \in B} \min_{a \in A} \{f(x_0, a, b) \cdot \nu\} \leq -\delta$$

for all exterior normal vectors ν at $x \in B(x_0, r) \cap \partial\mathcal{T}$. Then, for $\eta < \delta$, there exists $\varepsilon > 0$ such that $T(x) \leq d(x)/\eta$ for all $x \in B(x_0, \varepsilon) \setminus \mathcal{T}$. Moreover, T is locally Lipschitz continuous in a neighborhood of x_0 .

Proof. Let $\eta < \delta$. We define the set $S := \text{cl} \{ \nu : \nu \text{ is exterior normal at } x \in B(x_0, r) \cap \partial\mathcal{T} \}$, which is compact. Then, for all $\nu \in S$, we have $F(x_0, \nu) < -\eta$. We can therefore find $\varepsilon(\nu) > 0$ such that

$$(3.2) \quad F(x, \mu) < -\eta$$

for all $x \in B(x_0, \varepsilon(\nu))$ and $\mu \in S \cap B(\nu, \varepsilon(\nu))$. Then, if we extract a finite subcovering of S from the open covering $\{B(\nu, \varepsilon(\nu)) : \nu \in S\}$, determined by ν_1, \dots, ν_h , and we set $2\varepsilon := \min \{\varepsilon(\nu_i)\} \wedge r$, then (3.2) holds for all $x \in B(x_0, 2\varepsilon)$ and $\mu \in S$. Now let $x \in B(x_0, \varepsilon) \setminus \mathcal{T}$ and $p \in D^-d(x)$; by Proposition 1.6, $p = (x - z)/|x - z|$, where $z \in \partial\mathcal{T}$ is the unique point in \mathcal{T} such that $d(x) = |x - z|$ and thus $p \in S$. Therefore, we obtain

$$\min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot Dd(x)\} \geq \eta \quad \text{in } B(x_0, \varepsilon) \setminus \mathcal{T}$$

in the viscosity sense. If we now apply Theorem 3.6 to the function d , we obtain the requested estimate. As for the Lipschitz continuity of T , this is a standard result whenever the estimate with the distance function holds (see, e.g., [BS1] or [S1]). \square

Remark 3.8. Condition (3.1) is established for a general closed set \mathcal{T} . It is not difficult to verify that if \mathcal{T} is a closed C^2 manifold, then the set of exterior normal vectors at x_0 in the sense of Definition 1.5 coincides with the set of unit vectors of the normal space to \mathcal{T} at x_0 , which we indicate by $N_{x_0}(\mathcal{T})$. In this case, by the regularity of \mathcal{T} condition (3.1) is equivalent to

$$F(x_0, n) < 0 \quad \text{for all } n \in N_{x_0}(\mathcal{T}).$$

An analogous statement holds if \mathcal{T} is a C^2 manifold with boundary, but, in this case, if $x_0 \in \partial_{\text{rel}}(\mathcal{T})$, the set to be taken into account is $\{n \in N_{x_0}(\partial_{\text{rel}}(\mathcal{T})) : n \cdot n(x_0) \leq 0\}$, where $n(x_0)$ is the inner normal vector to \mathcal{T} at x_0 in the sense of manifolds.

If \mathcal{T} is a C^1 manifold (with boundary), the previous condition still gives the requested regularity of T , but, in this case, according to Definition 1.5, we could find points $x \in \partial\mathcal{T}$ at which the set of exterior normal vectors is empty; therefore, in general, we can only state that it is contained in the set of unit vectors of $N_{x_0}(\mathcal{T})$ (respectively of $\{n \in N_{x_0}(\partial_{\text{rel}}(\mathcal{T})) : n \cdot n(x_0) \leq 0\}$ if $x_0 \in \partial_{\text{rel}}(\mathcal{T})$). Corollary 3.7 generalizes to differential games and to general closed targets a result of local Lipschitz continuity of the minimum time function in control theory, which was proved by the author in [S1] by using different techniques for the case where \mathcal{T} is a C^1 manifold with boundary and the classical result of Petrov [P], in which $\mathcal{T} = \{0\}$. We should also note that (3.1) is a very general sufficient condition for the local Lipschitz continuity of T around x_0 . In fact, if \mathcal{T} is the closure of an open set, (3.1) is also necessary (see [BS1]).

In the theory of geometric local controllability, many sufficient conditions for the local Lipschitz and Hölder continuity of the minimum time function have been studied for the case where $\mathcal{T} = \{0\}$ or is a smooth manifold (see [S1], Stefani [St], and the references therein). Such conditions usually concern the Lie algebra generated by the vector field f on $\partial\mathcal{T}$. The next result gives a sufficient condition for local Hölder continuity of T in differential games, uses partial differential equation methods, and requires an assumption that concerns the field f in a neighborhood of \mathcal{T} . At the present time, we do not know which are the precise relationships between the two different approaches.

COROLLARY 3.9. *Let $x_0 \in \partial\mathcal{T}$, and assume that there exist $0 < \alpha < 1$, $C, r > 0$ such that*

$$(3.3) \quad \min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot Dd(x)\} \geq Cd^{1-\alpha}(x) \quad \text{in } B(x_0, r) \setminus \mathcal{T}$$

in the viscosity sense. Then there exists $r' \leq r$ such that $T(x) \leq d^\alpha(x)/(C\alpha)$ for all $x \in B(x_0, r') \setminus \mathcal{T}$. Moreover, T is locally Hölder continuous with exponent α in a neighborhood of x_0 .

Proof. We first note that condition (3.3) can be rewritten as was done in Corollary 3.7, as a condition on the exterior normal vectors at points in $B(x_0, 2r) \cap \partial\mathcal{T}$ because of the representation of the subdifferential of $d(\cdot)$ given in Proposition 1.6. We now observe that, by means of an easy calculation and the definition of subdifferential, we can recover the following equality between sets:

$$D^-d^\alpha(x) = \alpha d^{\alpha-1}(x) D^-d(x).$$

This fact and (3.3) immediately give that the function $U(x) := d^\alpha(x)/(\alpha C)$ satisfies the conditions of Theorem 2.9, from which we obtain the conclusion. \square

We now turn to the study of stability and local asymptotical terminability of a pursuit game and first prove a lemma that relates the problem of stability to the study of particular level sets.

LEMMA 3.10. *Let $h: \mathbf{R}^N \setminus \mathcal{T} \rightarrow \mathbf{R}$ be a positive continuous function and*

$$W(x) := \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} \int_0^{t_x} h(y_x(t)) dt,$$

then, for $x \in \mathbf{R}^N \setminus \mathcal{T}$ each $\varepsilon > 0$, there is $\alpha \in \Delta$ such that, for all $b \in \mathcal{B}$, we have $y_x(t) \in \{z: W(z) < W(x) + \varepsilon\}$ for all $t \in [0, t_x]$.

Proof. Let $\varepsilon > 0$, then by definition we can choose $\alpha \in \Delta$ such that

$$\int_0^{t_x} h(y_x(s)) ds < W(x) + \varepsilon \quad \text{for all } b \in \mathcal{B}.$$

Let $\bar{b} \in \mathcal{B}$ and $t \in (0, t_x)$, then, with obvious definitions of controls and strategies,

$$\int_0^t h(y_x(s)) ds + \int_0^{t_{y_x(t)}} h(y_{y_x(t)}(s)) ds < W(x) + \varepsilon$$

and then immediately $W(y_x(t)) < W(x) + \varepsilon$. \square

THEOREM 3.11. *Let $r > 0$ and $\Omega \supset B(\mathcal{T}, r)$ be an open set. Suppose that there exists a function $U \in C(\overline{\Omega \setminus \mathcal{T}})$, bounded below, which satisfies in the viscosity sense*

$$\min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot DU(x)\} \geq \mu(d(x)) \quad \text{in } \Omega \setminus \mathcal{T},$$

$$U = U_0 \quad \text{on } \partial\Omega,$$

where μ is a continuous, nonnegative real function, $\mu(t) = 0$ only if $t = 0$, and $\liminf_{t \rightarrow +\infty} \mu(t) > 0$. Assume that $U(x) < U_0$ if $x \in \Omega \setminus \mathcal{T}$ and that $|f| \leq M$ in $\partial\mathcal{T} \times A \times B$, then the game is locally asymptotically terminable. If, moreover, the level sets $\{x: U(x) < \varepsilon\}$ are a basis of the neighborhoods of \mathcal{T} (therefore, U is constant on $\partial\mathcal{T}$), then the game is also stable.

Proof. First, we assume that $U \geq 0$; otherwise, we will consider the function $U - \inf_{\Omega \setminus \mathcal{T}} U$. Let $x_0 \in \Omega \setminus \mathcal{T}$ and $\varepsilon < d(x_0) \wedge r$ such that $\Omega \supset B(\mathcal{T}, \varepsilon)$. We define t_x^ε as

the first exit time from $\mathbf{R}^N \setminus \overline{B(\mathcal{T}, \varepsilon)}$. Then U satisfies in the viscosity sense the following Dirichlet problem, where g_ε denotes the restriction of U on $\partial B(\mathcal{T}, \varepsilon)$:

$$\min_{b \in B} \max_{a \in A} \{-f(x, a, b) \cdot DU(x) - \mu(d(x))\} \geq 0 \quad \text{in } \Omega \setminus B(\mathcal{T}, \varepsilon),$$

$$U = g_\varepsilon \quad \text{on } \partial B(\mathcal{T}, \varepsilon),$$

$$U = U_0 \quad \text{on } \partial \Omega.$$

The usual solution of the previous problem is the value function

$$V_\varepsilon(x) := \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} \left\{ \int_0^{t_x^\varepsilon} \mu(d(y_x(t))) dt + g_\varepsilon(y_x(t_x^\varepsilon)) \right\}.$$

Therefore, since $\mu(d(x)) \geq \inf_{t \geq \varepsilon} \mu(t) > 0$ in $\Omega \setminus B(\mathcal{T}, \varepsilon)$, we can apply Theorem 2.5 and state that $V_\varepsilon(x) \leq U(x)$ for all $x \in \Omega \setminus B(\mathcal{T}, \varepsilon)$.

Now we set $\varepsilon_j := \varepsilon/2^j$, $E_0 := \Omega \setminus B(\mathcal{T}, \varepsilon)$ and $E_j := \overline{B(\mathcal{T}, \varepsilon_{j-1})} \setminus B(\mathcal{T}, \varepsilon_j)$ for $j \in \mathbf{N}$. Let $\eta > 0$ and observe that, by the definition of V_{ε_j} and the inequality $V_{\varepsilon_j} \leq U$, for all $x \in E_j$ there exists $\alpha_x \in \Delta$ such that

$$\int_0^{t_x^{\varepsilon_j}} \mu(d(y(t))) dt + U(y_x(t_x^{\varepsilon_j})) \leq U(x) + \eta/2^{j+1} \quad \text{for all } b \in \mathcal{B}.$$

We observe that this implies that $t_x^{\varepsilon_j} < +\infty$. Therefore, we choose $\bar{b} \in \mathcal{B}$ and denote $\alpha_0 = \alpha_{x_0}$, $b_0 = \bar{b}$, $x_1 = y(t_0)$, $t_0 = t_{x_0}^\varepsilon$. Then we have

$$\int_0^{t_0} \mu(d(y_{x_0}(t))) dt + U(x_1) \leq U(x_0) + \eta/2, \quad x_1 \in E_1.$$

In the general case, let $\sigma_{j-1} = t_0 + \dots, t_{j-1}$, $\alpha_j = \alpha_{x_j}$, $b_j(t) = \bar{b}(t + \sigma_{j-1})$, $t_j = t_{x_j}^{\varepsilon_j}$; then

$$(3.4) \quad \int_0^{t_j} \mu(d(y_{x_j}(t))) dt + U(x_{j+1}) \leq U(x_j) + \eta/2^{j+1}, \quad x_{j+1} \in E_{j+1}.$$

Thus, if we define

$$\bar{a}(t) := \begin{cases} \alpha_0[b_0](t) & \text{if } t \in [0, \sigma_0], \\ \alpha_1[b_1](t - \sigma_0) & \text{if } t \in [\sigma_0, \sigma_1] \\ \dots \end{cases}$$

and the nonanticipating strategy $\bar{\alpha}: \mathcal{B} \rightarrow \mathcal{A}$ by means of the position $\bar{\alpha}[\bar{b}] = \bar{a}$, then, by applying recursively (3.4) and since U is nonnegative, we have that, for such $\bar{\alpha}$ and all $b \in \mathcal{B}$,

$$\liminf_{t \rightarrow t_{x_0}(\bar{\alpha}[b], b)} d(y(t)) = 0 \quad \text{and} \quad \int_0^{t_{x_0}} \mu(d(y(t))) dt \leq U(x_0) + \eta.$$

This implies, on the one hand, the estimate

$$(3.5) \quad \inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} \int_0^{t_{x_0}} \mu(d(y(t))) dt \leq U(x_0) \quad \text{for all } x_0 \in \Omega \setminus \mathcal{T}$$

and, on the other hand, that the game is locally asymptotically terminable. In fact, if $t_{x_0} = +\infty$ and

$$\limsup_{t \rightarrow t_{x_0}(\bar{\alpha}[b], b)} d(y(t)) = \gamma > 0,$$

then, by using the notations above, for $\varepsilon_j < \gamma$ we can find an increasing sequence of positive real numbers $\{s_n: n \in \mathbb{N}\}$, such that, setting $y_n = y(s_n)$, we have $d(y_{2n}) = \varepsilon_j$, $d(y_{2n+1}) = \varepsilon_{j+1}$, and $\varepsilon_{j+1} < d(y(t)) < \varepsilon_j$ for $t \in (s_{2n}, s_{2n+1})$. Therefore, by the assumptions on the vector field, we obtain

$$\varepsilon_{j+1} \leq |y_{2n} - y_{2n+1}| \leq \int_{s_{2n}}^{s_{2n+1}} |f(y)| dt \leq (M + L\varepsilon_{j+1})|s_{2n} - s_{2n+1}|$$

and then

$$\begin{aligned} U(x_0) + \eta &\geq \int_0^{t_{x_0}} \mu(d(x)) dx \geq \sum_n \int_{s_{2n}}^{s_{2n+1}} \mu(d(y)) dt \\ &\geq \sum_n \inf_{t \geq \varepsilon_{j+1}} \mu(t)|s_{2n} - s_{2n+1}| \geq \sum_n \inf_{t \geq \varepsilon_{j+1}} \mu(t)\varepsilon_{j+1}/(M + L) = +\infty, \end{aligned}$$

which gives a contradiction. The last claim is a consequence of (3.5) and Lemma 3.10. \square

Remark 3.12. Note that inequality (3.5) in the proof of Theorem 3.11 holds without the supplementary assumption on the vector field. Now we extend to differential games the definition of L^p stability, which was introduced by Strauss [Sr] for dynamical systems and we state that a game is locally L^p terminable if there exists an open set $\Omega \in \mathcal{T}$ such that

$$\inf_{\alpha \in \Delta} \sup_{b \in \mathcal{B}} \int_0^{t_x} d^p(y_x(t; \alpha[b], b)) dt < +\infty \quad \text{for all } x \in \Omega \setminus \mathcal{T}.$$

By (3.5) we therefore obtain that, if $\mu(s) = cs^p$ for $c, p > 0$, then the game is L^p terminable.

Estimate (3.5) also holds independent interest. It can be seen as a comparison result between the value function of a generalized pursuit-evasion game whose running cost becomes zero on the target and a supersolution of the corresponding Isaacs equation. In fact, if the supersolution U also satisfies $U = 0$ on $\partial\mathcal{T}$, then (3.4) implies (3.5) if either U is nonnegative or it takes up the boundary data on $\partial\mathcal{T}$ uniformly. It is easy to verify that, in the general case of § 2, we need the following assumption on the sign of the running cost:

$$h(x, a, b) \geq C_\varepsilon > 0 \quad \text{for all } \varepsilon > 0, x \in \mathbf{R}^N \setminus B(\mathcal{T}, \varepsilon), a, b.$$

The corresponding statement holds for subsolutions as well.

Remark 3.13. All the previous results also provide sufficient conditions of global and global asymptotic terminability when $\Omega = \mathbf{R}^N$. In this case, the conditions at the points of $\partial\Omega$ obviously disappear. We finally observe that the choice of the distance function in the statement is only for simplicity.

3.2. Evasion problem. We now apply a similar method to the study of the counterpart of the pursuit problem.

THEOREM 3.14. *Let $\Omega \supset \mathcal{T}$ be an open set. Assume that there exists $U \in C(\overline{\Omega \setminus \mathcal{T}})$, which satisfies in the viscosity sense*

$$\max_{a \in A} \min_{b \in B} \{-f(x, a, b) \cdot DU(x)\} \leq 0 \quad \text{in } \Omega \setminus \mathcal{T}.$$

If, moreover, $U(x) = 0$ for $x \in \partial\mathcal{T}$, $U > 0$ in $\Omega \setminus \mathcal{T}$, and $c := \inf_{\partial\Omega} U > 0$, then the game is evadable.

Proof. We consider the set $\Omega' := \{x \in \Omega : U(x) < c\}$ and observe that $U(x) = c$ for all $x \in \partial\Omega'$. We first consider the case where $x_0 \in \Omega'$, therefore, we can choose suitable ε, σ such that $0 \leq \varepsilon < U(x_0) < \sigma \leq c$, and prove that the game is evadable at x_0 . We define $\tilde{U} := U - \varepsilon$ and observe that if we set, for $\eta > 0$, $\Omega_\eta := \{x \in \Omega : U(x) < \eta\}$, then \tilde{U} satisfies in the viscosity sense

$$(3.6) \quad \begin{aligned} \max_{a \in A} \min_{b \in B} \{-f(x, a, b) \cdot D\tilde{U}(x)\} &\leq 0 \quad \text{in } \Omega_\sigma \setminus \Omega_\varepsilon, \\ \tilde{U} &= 0 \quad \text{on } \partial\Omega_\varepsilon, \quad \tilde{U}(x) = \sigma - \varepsilon \quad \text{on } \partial\Omega_\sigma, \end{aligned}$$

and $\tilde{U} < \sigma - \varepsilon$ in $\Omega_\sigma \setminus \Omega_\varepsilon$. By a simple calculation, it is easy to verify that the differential inequality in system (3.6) is also satisfied by $\tilde{u} = \psi(\tilde{U})$ (ψ is defined in (1.5)) with boundary conditions $\tilde{u} = 0$ on $\partial\Omega_\varepsilon$, $\tilde{u}(x) = \psi(\sigma - \varepsilon)$ on $\partial\Omega_\sigma$. To this end, it is enough to observe that, if $\tilde{U} - \varphi$ attains a maximum point at y and $\tilde{U}(y) = \varphi(y)$, since

$$\psi(\tilde{U} - \varphi) = \exp(\varphi)(\psi(\tilde{U}) - \psi(\varphi)),$$

then also $\tilde{u} - \psi(\varphi)$ attains a maximum point at y and, moreover, $D\varphi = \exp(\varphi)D\psi(\varphi)$.

Now we set $u(x) := \tilde{u}(x)/\psi(\sigma - \varepsilon)$ and observe that $u < 1$ in $\Omega_\sigma \setminus \Omega_\varepsilon$; therefore, by (3.6) u satisfies the following Dirichlet problem in the viscosity sense:

$$\begin{aligned} u(x) + \mathcal{H}(x, u(x), Du(x)) &= u(x) - 1 + \max_{a \in A} \min_{b \in B} \{-f(x, a, b) \cdot Du(x)\} \leq 0 \quad \text{in } \Omega_\sigma \setminus \Omega_\varepsilon, \\ u &= 0 \quad \text{on } \partial\Omega_\varepsilon, \quad u = 1 \quad \text{on } \partial\Omega_\sigma. \end{aligned}$$

Therefore, if $t_x^\varepsilon(a, b)$ denotes the first exit time from $\mathbf{R}^N \setminus \overline{\Omega_\varepsilon}$, \tilde{T}_ε is the corresponding upper capture time, and $\tilde{v}_\varepsilon := \psi(\tilde{T}_\varepsilon)$, then, by the proof of Theorem 2.3, we deduce that

$$(3.7) \quad \psi(U(x) - \varepsilon)/\psi(\sigma - \varepsilon) = u(x) \leq \tilde{v}_\varepsilon(x) \quad \text{in } \Omega_\sigma \setminus \Omega_\varepsilon.$$

If we compute (3.7) at x_0 and let $\sigma \rightarrow U(x_0)$ from above, since $\sigma > U(x_0)$ were arbitrary, then we obtain $\tilde{T}_\varepsilon(x_0) = +\infty$ (note that the limit and the statement do not depend on the choice of $\varepsilon < U(x_0)$). This is not enough to conclude, but with a construction similar to the one in the proof of Theorem 3.11 it is easy to define a strategy $\bar{\beta} : \mathcal{A} \rightarrow \mathcal{B}$ such that

$$t_{x_0}(\bar{a}, \bar{\beta}[\bar{a}]) = +\infty \quad \text{for each choice of } \bar{a},$$

which means that $y_x(t; \bar{a}, \bar{\beta}[\bar{a}]) \in \mathbf{R}^N \setminus \mathcal{T}$ for all $t \geq 0$, and thus the game is evadable at x_0 . In fact, β can be chosen in such a way that the trajectory $y(t; \bar{a}, \bar{\beta}[\bar{a}])$ remains outside Ω_ρ for all fixed $\rho < U(x_0)$, and this will be used in the next result. With the same technique, it is now easy to treat the case where $x_0 \in \mathbf{R}^N \setminus \Omega'$. \square

As a consequence of the last remark in the previous proof, we state the following result.

COROLLARY 3.15. *In the framework of the previous theorem, assume, moreover, that there exists a strictly increasing function $\mu : \mathbf{R}_+ \rightarrow \mathbf{R}_+$ such that $\mu(t) = 0$ only if $t = 0$ and satisfies*

$$U(x) \leq \mu(d(x)) \quad \text{for all } x \in \Omega \setminus \mathcal{T}.$$

Then the game is strictly evadable.

Proof. We need only observe that, for $\varepsilon > 0$, we have $B(\mathcal{T}, \mu^{-1}(\varepsilon)) \subset \Omega_\varepsilon$, and we apply the proof of Theorem 3.14, in particular, the last remark. \square

As in the pursuit problem, the first candidate function to fulfill the hypotheses of Theorem 3.14 is the distance function. In view of the characterization of the super-differential of $d(\cdot)$ in Proposition 1.6, we state the following.

COROLLARY 3.16. *Assume one of the three following conditions:*

(i) *There exists $r > 0$ such that*

$$\tilde{F}(x, \nu) := \min_{a \in A} \max_{b \in B} \{f(x, a, b) \cdot \nu\} \geq 0 \quad \text{for all } x \in B(\mathcal{T}, r) \setminus \mathcal{T}$$

and for all vectors $\nu \in \text{co} \{(x - z)/|x - z| : z \in \partial\mathcal{T}, d(x) = |x - z|\}$ (co X indicates the convex hull of the set X).

(ii) *There exist $\delta, r > 0$ such that $d(\cdot)$ is differentiable in $B(\mathcal{T}, r) \setminus \mathcal{T}$ and $\tilde{F}(x_0, \nu) > \delta$, for all $x_0 \in \partial\mathcal{T}$ and all the exterior normal vectors ν at $x \in B(x_0, r) \cap \partial\mathcal{T}$.*

(iii) *Let $\partial\mathcal{T}$ be bounded and, for all $x_0 \in \partial\mathcal{T}$, there exists $r_0(x_0) > 0$ such that $\tilde{F}(x_0, \nu) > 0$ for all the exterior normal vectors ν at $x \in B(x_0, r_0) \cap \partial\mathcal{T}$.*

Then the game is strictly evadable.

Proof. Assumption (i) implies that $d(\cdot)$ satisfies the differential inequality

$$\max_{a \in A} \min_{b \in B} \{-f(x, a, b) \cdot Dd(x)\} \leq 0 \quad \text{for all } x \in B(\mathcal{T}, r) \setminus \mathcal{T}$$

in the viscosity sense. Now it is enough to apply Corollary 3.15.

Cases (ii) and (iii) are just modifications, using, in particular, the compactness arguments of the proof of Corollary 3.7. \square

REFERENCES

- [A] J. P. AUBIN, *A survey of viability theory*, SIAM J. Control Optim., 4 (1990), pp. 749–778.
- [AC] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, New York, 1984.
- [AF] J. P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser-Verlag, Basel, Switzerland, 1990.
- [Ba] G. BARLES, *Remarques sur des resultats d'existence pour les equations de Hamilton–Jacobi du premier ordre*, Ann. Inst. H. Poincaré, 1 (1985), pp. 21–32.
- [BSY] N. P. BATHIA, G. P. SZEGO, AND J. A. YORKE, *A Liapunov characterization of attractors*, Boll. Un. Mat. Ital., 2 (1969), pp. 222–228.
- [B] J. M. BONY, *Principe du maximum, inégalité de Harnak et unicité du problème de Cauchy pour les opérateurs elliptiques dégénérés*, Ann. Inst. Fourier (Grenoble), 19 (1969), pp. 277–304.
- [BS1] M. BARDI AND P. SORAVIA, *Hamilton–Jacobi equations with singular boundary conditions on a free boundary and applications to differential games*, Trans. Amer. Math. Soc., 325 (1991), pp. 205–229.
- [BS2] ———, *Time optimal control, Lie brackets and Hamilton–Jacobi equations*, preprint, University of Padua, Padua, Italy, 1991.
- [BS3] ———, *A comparison result for Hamilton–Jacobi equations and applications to different games lacking controllability*, Funkcial. Ekvac., to appear.
- [Br] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential and Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussman, eds., Birkhäuser-Verlag, Basel, Switzerland, 1983.
- [C] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [CEL] M. C. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
- [CL] M. C. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
- [CN] M. C. CRANDALL AND R. NEWCOMB, *Viscosity solutions of Hamilton–Jacobi equations at the boundary*, Proc. Amer. Math. Soc., 94 (1985), pp. 283–290.
- [EI] L. C. EVANS AND H. ISHII, *Differential games and nonlinear first order PDE on bounded domains*, Manuscripta Math., 49 (1984), pp. 109–139.
- [EJ] L. C. EVANS AND M. R. JAMES, *The Hamilton–Jacobi–Bellman equation for time-optimal control*, SIAM J. Control Optim., 27 (1989), pp. 1477–1489.

- [ES] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 773-797.
- [EK1] R. J. ELLIOTT AND N. J. KALTON, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972).
- [EK2] ———, *Cauchy problems for certain Isaacs-Bellman equations and games of survival*, Trans. Amer. Math. Soc., 198 (1974), pp. 45-72.
- [Fl] W. H. FLEMING, *The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1961), pp. 102-116.
- [Fr] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971.
- [I1] H. ISHII, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369-384.
- [I2] ———, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 105-135.
- [LL] V. LAKSHMIKANTHAM AND S. LEELA, *Differential and Integral Inequalities*, Academic Press, New York, 1969.
- [LM] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [L1] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi equations*, Pitman, Boston, 1982.
- [L2] ———, *Existence results for first order Hamilton-Jacobi equations*, Ricerche Mat., 32 (1983), pp. 3-23.
- [LSO] P. L. LIONS AND P. E. SOUGANIDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations*, SIAM J. Control Optim., 23 (1985), pp. 566-583.
- [P] N. N. PETROV, *Controllability of autonomous systems*, Differentsial nye Uravneniya, 4 (1968), pp. 606-617.
- [Ro] R. T. ROCKAFELLAR, *The Theory of Subgradients and its Application to Problems of Optimization: Convex and Nonconvex Functions*, R & E, Heldermann-Verlag, Berlin, 1981.
- [R] E. ROXIN, *Axiomatic approach in differential games*, J. Optim. Theory Appl., 3 (1969), pp. 153-163.
- [Sn] H. M. SONER, *Optimal control problems with state-space constraints I and II*, SIAM J. Control Optim., 24 (1987), pp. 551-561, 1110-1122.
- [S1] P. SORAVIA, *Hölder continuity of the minimum time function for C^1 manifold targets*, J. Optim. Theory Appl., 75 (1992).
- [S2] ———, *The concept of value in differential games of survival and viscosity solutions of Hamilton-Jacobi equations*, Differential Integral Equations, to appear.
- [So] P. E. SOUGANIDIS, *A remark about viscosity solutions of Hamilton-Jacobi equations at the boundary*, Proc. Amer. Math. Soc., 96 (1986), pp. 323-329.
- [St] G. STEFANI, *Regularity properties of the minimum time function*, in Proc. Internat. Inst. for Applied Systems Analysis Conference, Sopron, Hungary, 1989.
- [Sr] A. STRAUSS, *Liapunov functions and L^p solutions of differential equations*, Trans. Amer. Math. Soc., 119 (1965), pp. 37-50.
- [V] P. P. VARAIYA, *On the existence of solutions to a differential game*, SIAM J. Control, 5 (1967), pp. 153-162.
- [Y1] J. YONG, *On differential pursuit games*, SIAM J. Control Optim., 26 (1988), pp. 478-495.
- [Y2] ———, *On differential evasion games*, SIAM J. Control Optim., 26 (1988), pp. 1-22.
- [Yo] J. A. YORKE, *Differential inequalities and non-Lipschitz scalar functions*, Math. Systems Theory, 4 (1970), pp. 140-153.

AN OPTIMAL CONTROL PROBLEM WITH A MIXED COST FUNCTION*

V. I. KOROBOV†, V. I. KRUTIN†, AND G. M. SKLYAR†

Abstract. The main objective of this paper is to find an optimal positional control in feedback form for a linear autonomous system with a mixed cost functional. The method of controllability functions is used. The relationship between this problem and an optimal stabilizability problem is also investigated, and several examples are presented.

Key words. linear system, mixed cost function, optimal positional control, controllability function, Riccati equation

AMS(MOS) subject classifications. 49C20, 34H05

Introduction. Consider the optimal control problem of the system

$$(0.1) \quad \dot{x} = Ax + Bu, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^r, \quad \text{rg}(B, AB, \dots, A^{n-1}B) = n,$$

$$(0.2) \quad x(0) = x_0, \quad x(T) = 0, \quad T \text{ arbitrary},$$

for which it is desired to minimize the functional

$$(0.3) \quad J(u) = k^2 T + \int_0^T [(Wx, x) + (Uu, u)] dt, \quad k > 0,$$

where W, U are symmetric matrices with $W \geq 0, U > 0$.

We refer to this functional as a mixed cost function. The main goal of our work is to find an optimal positional control for the formulated problem, i.e., to find the function $u = u(x)$, which depends only on the state variable x such that

(i) for any $x_0 \in \mathbb{R}^n$, the solution $x(t)$ of the Cauchy problem

$$(0.4) \quad \dot{x} = Ax + Bu(x), \quad x(0) = x_0$$

exists on some interval $[0, T(x_0))$ and is unique;

(ii) $x(t) \rightarrow 0$ as $t \rightarrow T(x_0)$;

(iii) the pair $(x(t), u(x(t)))$ is the solution of the optimal control problem (0.1)–(0.3).

In § 1 we consider the relationship between the optimal positional control problem and a traditional problem of finding the optimal control for (0.1)–(0.3) in the nonstationary linear feedback form. A form of the positional control $u(x)$ is given. It is shown that this control is not linear, as compared to the well-known case when $k = 0$ [1]–[3].

Therefore the investigation of the existence and uniqueness of the Cauchy problem (0.4) solution (§ 2) as defined in this article comprises an essential part of this work.

Property (ii) is obtained by the method of the controllability function, proposed in [4], on the basis of which different problems of nonoptimal admissible positional synthesis were solved; see, for example, [4], [5].

In § 3 optimality of the Cauchy problem (0.4) solution is proved (property (iii)). The case where $W = 0$ is considered separately, where we can also find a form of the optimal trajectories.

* Received by the editors August 14, 1989; accepted for publication (in revised form) July 18, 1991.

† Department of Differential Equations and Controls, Kharkov State University, 4 Freedom Square, 310077 Kharkov, Ukraine.

In the concluding sections, the relationship with the solution of the optimal stabilizability problem is established, i.e., with finding $u = u(x)$ such that $x(T) = 0$ and

$$(0.5) \quad \int_0^T [(Wx, x) + (Uu, u)] dt \rightarrow \inf.$$

It is shown that, when $k \rightarrow 0$ and $W > 0$, we have that $u_k(x) \rightarrow u(x)$, where $u_k(x)$ is the optimal positional control in problem (0.1)–(0.3) and $u(x)$ is the optimal positional control in problem (0.1), (0.5). This limited transition is also investigated in detail when $W = 0$.

1. Finding the positional control $u(x)$. It is a well-known fact that the solution of problem (0.1)–(0.3) for a fixed $T > 0$ exists and can be found in a nonstationary feedback form

$$(1.1) \quad u(x, t) = -U^{-1}B^*N^{-1}(T-t)x,$$

where $N(t)$ is the solution of the Cauchy problem for the matrix Riccati equation

$$(1.2) \quad \dot{N} = -AN - NA^* + BU^{-1}B^* - NWN, \quad N(0) = 0.$$

This will be helpful to us throughout the paper.

The optimal value of the functional (0.3) here is

$$J^*(T) = k^2T + (N^{-1}(T)x_0, x_0).$$

It is essential here that the Cauchy problem (1.2) has its only solution in the class of positive definite matrices for $t > 0$ [6]–[8] and is an analytic function of $t > 0$, according to the Kovalevskaya theorem.

In the case when T is not fixed, we let

$$J^* = \inf_{T>0} J_T^* = \inf_{T>0} [k^2T + (N^{-1}(T)x_0, x_0)].$$

Let us designate, for $T > 0$ and $x \in \mathbb{R}^n$, that

$$(1.3) \quad \psi(T, x) = k^2T + (N^{-1}(T)x, x),$$

$$(1.4) \quad \begin{aligned} \Phi(T, x) &= \psi'_T(T, x) \\ &= k^2 + ([N^{-1}(T)A + A^*N^{-1}(T) - N^{-1}(T)BU^{-1}B^*N^{-1}(T) + W]x, x). \end{aligned}$$

Remark 1.1. At each $x \in \mathbb{R}^n \setminus \{0\}$, the function $\psi(T, x)$ achieves its minimum. In fact, the function $\psi(T, x)$ is continuous (moreover, this function is analytic on $(0, \infty) \times \mathbb{R}^n$). Since $\|N(T)\|^{-1}\|x\| < (N^{-1}(T)x, x)$ and $\lim_{T \rightarrow +0} \|N(T)\| = 0$, it follows that $\lim_{T \rightarrow +0} \psi(T, x) = +\infty$.

On the other hand, it is also clear that $\lim_{T \rightarrow +\infty} \psi(T, x) = +\infty$; so the statement is verified.

Let $I_x = \{T > 0: \psi(T, x) \leq \psi(t, x) \text{ for every } t > 0\}$, $x \in \mathbb{R}^n \setminus \{0\}$. It follows from Remark 1.1 that the set $I_x \neq \emptyset$, and it follows from analyticity of the function $\psi(t, x)$ that this set consists of a finite number of points. Moreover, if $T \in I_x$, then $\Phi(T, x) = 0$.

Thus we have the following remark.

Remark 1.2. Problem (0.1)–(0.3) has a solution for every $x_0 \in \mathbb{R}^n \setminus \{0\}$. Moreover, the set of optimal controls has the form (1.1), where $T \in I_{x_0}$.

Remark 1.2 gives the solution of problem (0.1)–(0.3) in a nonstationary linear feedback form.

Next, we proceed to the choice of the optimal positional control $u(x)$. According to the required property (iii) of $u(x)$, at each $x_0 \in \mathbb{R}^n \setminus \{0\}$, $u(x_0)$ must coincide with the value of some optimal control of the form (1.1) when $t = 0$; i.e.,

$$(1.5) \quad u(x_0) \in \mathbb{Q}_{x_0} = \{u_0; u_0 = -U^{-1}B^*N^{-1}(T)x_0, T \in I_{x_0}\}.$$

If at each x_0 the set \mathbb{Q}_{x_0} had precisely one point, then the choice of the positional control $u(x)$ is trivial. This control would evidently satisfy properties (i)–(iii), with the possible exception of the uniqueness of the solution of problem (0.4). Such a situation takes place, for example, in the case when $k = 0$ (linear-quadratic problem) or in the time optimal problem [9], where the programmed optimal control for each initial point is uniquely defined. However, in the problem considered here, for some x , the set I_x may contain more than one point (for an example, see § 4). Therefore the choice of the unique control $u(x) \in \mathbb{Q}_x$, $x \in \mathbb{R}^n \setminus \{0\}$ having properties (i)–(iii), presents serious difficulties. Moreover, for any such choice, the control $u(x)$ will be discontinuous.

Later in this paper, it is shown that the control $u(x)$ defined by

$$(1.6) \quad u(x) = -U^{-1}B^*N^{-1}(\theta(x))x, \quad x \neq 0,$$

$$(1.7) \quad \theta(x) = \min I_x,$$

is the optimal positional control.

We refer to the function that is defined for $x \neq 0$ by equality (1.6) and defined for $x = 0$ by (1.6) with $\theta(0) = 0$ as the controllability function. At each $x \neq 0$, the value of this function coincides with the minimal positive root of the equation $\Phi(\theta, x) = 0$, at which the function $\psi(\theta, x)$ attains its global minimum.

We conclude this section by establishing some properties of the controllability function.

Property 1. From $\theta(x) \rightarrow 0$, it follows that $x \rightarrow 0$.

In fact,

$$(1.8) \quad \|N(\theta(x))\|^{-1}\|x\|^2 \leq \psi(\theta(x), x) = k^2\theta + (N^{-1}(\theta(x))x, x).$$

On the other hand,

$$(1.9) \quad \psi(\theta, x) \leq \psi(1, x) = k^2 + (N^{-1}(1)x, x) \leq k^2 + \|N^{-1}(1)\|\|x\|^2.$$

From (1.8) and (1.9), we obtain that

$$(1.10) \quad \|x\|^2(\|N(\theta(x))\|^{-1} - \|N^{-1}(1)\|) \leq k^2.$$

Let ε be any positive number such that

$$(1.11) \quad 1 - \|N^{-1}(1)\|\varepsilon > 0.$$

Choose $\delta > 0$ such that, if $0 < \theta(x) < \delta$, then $\|N(\theta(x))\| < \varepsilon$; that is, $\|N(\theta(x))\|^{-1} > 1/\varepsilon$. Then, when $0 < \theta(x) < \delta$, because of (1.10) and (1.11),

$$\|x\|^2(1/\varepsilon - \|N^{-1}(1)\|) \leq k^2, \quad \|x\|^2 \leq \varepsilon k^2 / (1 - \|N^{-1}(1)\|\varepsilon)$$

are fulfilled; hence Property 1 is obtained.

Property 2. When $\theta(x) \rightarrow 0$, then $\psi(\theta(x), x) \rightarrow 0$.

Let ε be any positive number and choose $\varepsilon > \delta > 0$ so small that when $\theta(x) < \delta$, $(N^{-1}(\varepsilon)x, x) < \varepsilon$ (this is possible because of Property 1). Then $\psi(\theta(x), x) \leq \psi(\varepsilon, x) = (k^2 + 1)\varepsilon$, which proves Property 2.

2. Existence and uniqueness of the Cauchy problem solution.

DEFINITION 2.1. We call the function $x(t)$ that satisfies almost everywhere the relationship

$$(2.1) \quad \dot{x} = Ax - BU^{-1}B^*N^{-1}(\theta(x))x$$

and is such that the function $\theta(x(t))$ is absolutely continuous on t , as the solution of the differential equation (0.1) with control (1.6). As is seen from the definition, the solution $x(t)$ is a continuously differentiable function.

Remark 2.1. Consider the system

$$(2.2) \quad \begin{aligned} \dot{x} &= Ax - BU^{-1}B^*N^{-1}(\sigma)x, \\ \dot{\sigma} &= f(t), \quad x(0) = x_0, \quad \sigma(0) = \theta(x_0). \end{aligned}$$

As is seen from Definition 2.1, the function $x(t)$ is the solution of (2.1) if and only if there exists $f(t)$ from the class of locally summable functions such that the pair $(x(t), \sigma(t))$ of the solutions of system (2.2) satisfies the equation

$$(2.3) \quad \begin{aligned} \Phi(\sigma(t), x(t)) &= k^2 + ([N^{-1}(\sigma(t))A + A^*N^{-1}(\sigma(t)) \\ &\quad - N^{-1}(\sigma(t))BU^{-1}B^*N^{-1}(\sigma(t)) + W]x(t), x(t)) = 0, \end{aligned}$$

and, for every t , $\sigma(t) = \theta(x(t))$.

THEOREM 2.1. For any initial condition $x_0 \in \mathbb{R}^n$, the solution $x(t)$ of (2.1), in the sense of Definition 2.1, exists and is unique on the interval $[0, \theta(x_0))$, and $\dot{\theta}(x(t)) = -1$ for every $t \in [0, \theta(x_0))$.

The proof uses the following lemma.

LEMMA 2.1. The function $\Phi(\theta, x)$ determined by (1.4) satisfies the equation

$$(2.4) \quad \Phi_\theta = (\Phi_x, Ax - BU^{-1}B^*N^{-1}(\theta)x).$$

(Here and elsewhere, a subscript variable signifies the corresponding partial derivative.)

Proof. Equation (1.4) gives

$$\Phi(\theta, x) = k^2 + 2(Ax, N^{-1}(\theta)x) - (N^{-1}(\theta)x, BU^{-1}B^*N^{-1}(\theta)x) + (Wx, x).$$

Since $N(\theta)$ satisfies the differential equation (1.2), the inverse matrix satisfies the equation

$$(2.5) \quad N_\theta^{-1} = N^{-1}A + A^*N^{-1} - N^{-1}BU^{-1}B^*N^{-1} + W.$$

By using (2.5), we obtain that

$$\begin{aligned} \Phi_\theta &= 2(Ax, [N^{-1}(\theta)A + A^*N^{-1}(\theta) - N^{-1}(\theta)BU^{-1}B^*N^{-1}(\theta) + W]x) \\ &\quad - 2([N^{-1}(\theta)A + A^*N^{-1}(\theta) - N^{-1}(\theta)BU^{-1}B^*N^{-1}(\theta) + W] \\ &\quad \cdot x, BU^{-1}B^*N^{-1}(\theta)x) \\ &= -(2[N^{-1}(\theta)A + A^*N^{-1}(\theta) - N^{-1}(\theta)BU^{-1}B^*N^{-1}(\theta) + W] \\ &\quad \cdot x, Ax - BU^{-1}B^*N^{-1}(\theta)x) \\ &= (\Phi_x, Ax - BU^{-1}B^*N^{-1}(\theta)x), \end{aligned}$$

which proves the lemma.

Proof of Theorem 2.1. We first prove existence of a solution of (2.1).

Assume that in (2.2)

$$(2.6) \quad f(t) = -1, \quad t \in [0, \theta(x_0)).$$

Let $(x(t), \sigma(t))$ be the solution of (2.2), with $f(t)$ given by (2.6). Then, according to Lemma 2.1,

$$\frac{d}{dt} \Phi(\sigma(t), x(t)) = -\Phi_\theta + (\Phi_x, Ax(t) - BU^{-1}B^*N^{-1}(\sigma(t)), x(t)) = 0.$$

Since, at the initial instant $\sigma(0) = \theta(x_0)$, it follows that

$$\Phi(\sigma(t), x(t)) \equiv 0 \quad \text{for every } t \in [0, \theta(x_0)).$$

We claim that, for every $t \in [0, \theta(x_0))$, $\sigma(t) = \theta(x(t))$, where $\sigma(t)$ is the minimal root of the equation

$$(2.7) \quad \Phi(\theta, x(t)) = 0,$$

at which $\psi(\theta, x(t))$ attains its global minimum.

To see this, consider the function of two variables

$$G(\theta, t) = \psi(\theta, x(t)), \quad t \in [0, \theta(x_0)).$$

LEMMA 2.2. *For every fixed $\theta_1 > 0$, it holds that*

$$\dot{G}(\theta_1 - t, t) \geq \dot{G}(\theta(x_0) - t, t) = G(\sigma(t), t), \quad t \in [0, \min(\theta_1, \theta(x_0))].$$

Proof. By using (1.4), (2.5), and (2.6), we have that

$$(2.8) \quad \begin{aligned} \dot{G}(\sigma(t), t) &= (N^{-1}(\sigma(t))x(t), Ax(t) - BU^{-1}B^*N^{-1}(\sigma(t))x(t)) \\ &\quad + (Ax(t) - BU^{-1}B^*N^{-1}(\sigma(t))x(t), N^{-1}(\sigma(t))x(t)) - \Phi(\sigma(t), x(t)) \\ &= -k^2 - (Wx(t), x(t)) - (N^{-1}(\sigma(t))BU^{-1}B^*N^{-1}(\sigma(t))x(t), x(t)), \end{aligned}$$

as well as

$$(2.9) \quad \begin{aligned} \dot{G}(\theta_1 - t, t) &= -(N^{-1}(\sigma(t))BU^{-1}B^*N^{-1}(\theta_1 - t)x(t), x(t)) - k^2 \\ &\quad - ([N^{-1}(\theta_1 - t)A + A^*N^{-1}(\theta_1 - t) \\ &\quad - N^{-1}(\theta_1 - t)BU^{-1}B^*N^{-1}(\theta_1 - t) + W]x(t), x(t)) \\ &\quad + (A^*N^{-1}(\theta_1 - t)x(t), x(t)) - (N^{-1}(\theta_1 - t)Ax(t), x(t)) \\ &\quad - (N^{-1}(\theta_1 - t)BU^{-1}B^*N^{-1}(\sigma(t))x(t), x(t)). \end{aligned}$$

By subtracting (2.8) from (2.9), we obtain that

$$\begin{aligned} \dot{G}(\theta_1 - t, t) - \dot{G}(\sigma(t), t) &= (N^{-1}(\sigma(t))BU^{-1}B^*N^{-1}(\sigma(t))x(t), x(t)) \\ &\quad + (N^{-1}(\theta_1 - t)BU^{-1}B^*N^{-1}(\theta_1 - t)x(t), x(t)) \\ &\quad - (N^{-1}(\theta_1 - t)BU^{-1}B^*N^{-1}(\sigma(t))x(t), x(t)) \\ &\quad - (N^{-1}(\sigma(t))BU^{-1}B^*N^{-1}(\theta_1 - t)x(t), x(t)) \\ &= \|U^{-1/2}B^*(N^{-1}(\theta_1 - t) - N^{-1}(\sigma(t)))x(t)\|^2 \\ &\geq 0; \end{aligned}$$

so the lemma is proved.

Continuing with the proof of Theorem 2.1, we let $t_1 \in [0, \theta(x_0))$ and $y \in (0, +\infty)$. Then, by denoting $\theta_1 = y + t_1$, we obtain the following inequality from Lemma 2.2:

$$\dot{G}(\theta_1 - t, t) \geq \dot{G}(\sigma(t), t) \quad \text{for every } t \in [0, \min(\theta_1, \theta(x_0))].$$

Integration of this inequality from 0 to t_1 yields

$$G(\theta_1 - t_1, t_1) - G(\theta_1, 0) \geq G(\sigma(t_1), t_1) - G(\theta(x_0), 0),$$

which implies that

$$\begin{aligned} G(\theta_1 - t_1, t_1) - G(\sigma(t_1), t_1) &= \psi(y, x(t_1)) - \psi(\sigma(t_1), x(t_1)) \\ &\geq G(\theta_1, 0) - G(\theta(x_0), 0) \\ &= \psi(\theta_1, x_0) - G\psi(\theta(x_0), x_0) \\ &\geq 0. \end{aligned}$$

If $y < \sigma(t_1)$, then $\theta_1 < \theta(x_0)$ and $\psi(\theta_1, x_0) - \psi(\theta(x_0), x_0) > 0$ because $\theta(x_0)$ is the smallest value of θ for which the function $\psi(\theta, x_0)$ achieves its minimum. Therefore $\psi(y, x(t_1)) - \psi(\sigma(t_1), t_1) > 0$.

We infer that $\sigma(t)$ is the minimal root of (2.3), at which $\psi(\theta, x(t))$ attains its global minimum for every $t \in [0, \theta(x_0)]$; that is, $\sigma(t) = \theta(x(t))$. Consequently, the existence of the solution of (2.1) with control (1.6), in the sense of Definition 2.1, is proved.

Next, we show the uniqueness of the solution of (2.1). Partition $\mathbb{R}^n \setminus \{0\}$ as follows: $\mathbb{R}^n \setminus \{0\} = \mathbb{M}_1 \cup \mathbb{M}_2$, where

$$\mathbb{M}_1 = \{x \in \mathbb{R}^n \setminus \{0\} : \Phi'_\theta(\theta(x), x) \neq 0\}, \quad \mathbb{M}_2 = \{x \in \mathbb{R}^n \setminus \{0\} : \Phi'_\theta(\theta(x), x) = 0\}.$$

We consider separately the cases where $x_0 \in \mathbb{M}_1$ and $x_0 \in \mathbb{M}_2$, where x_0 is the initial condition.

Let $x_0 \in \mathbb{M}_1$. Then the equality $\Phi(\theta(x_0), x_0) = 0$ holds at the point x_0 , and $\Phi'_\theta(\theta(x_0), x_0) \neq 0$. Let $x(t)$ be an arbitrary solution of (2.1) with initial condition $x(0) = x_0$, in the sense of Definition 2.1, and consider the function $g(\theta, t) = \Phi(\theta, x(t))$. Then g is continuous and has continuous partial derivatives in θ and t in a certain neighbourhood of the point $(\theta(x_0), 0)$. Since $g_\theta(\theta(x_0), 0) \neq 0$, the implicit function theorem implies that there exists a unique absolutely continuous function $\sigma(t)$ such that $g(\sigma(t), t) = \Phi(\sigma(t), x(t)) = 0$, and, according to (2.4),

$$(2.10) \quad \dot{\sigma} = -(\Phi_x, Ax(t) - BU^{-1}B^*N^{-1}(\theta(x(t)))x(t))\Phi_\theta = -1.$$

Therefore $x(t)$ is the solution of system (2.2) when $f(t) = 1$, which proves uniqueness for any point $x_0 \in \mathbb{M}_1$.

The case where $x_0 \in \mathbb{M}_2$ requires a more delicate argument. By the definition of the function $\theta(x)$ and the analyticity of $\psi(\theta, x)$ in its first argument, there exists the finite number $n = 2k$, $k > 1$ such that

$$\frac{\partial^i \Phi(\theta(x_0), x_0)}{\partial \theta^i} = 0, \quad i = 1, \dots, n; \quad \frac{\partial^{n+1} \Phi(\theta(x_0), x_0)}{\partial \theta^{n+1}} \neq 0.$$

Let $x(t)$ be an arbitrary solution (2.1) with the initial condition $x(0) = x_0$ in the sense of Definition 2.1. Then there exists an interval $[0, \tau(x_0)]$, $\tau(x_0) > 0$ such that

$$(2.11) \quad \frac{\partial^{n+1} \Phi(\theta(x(t)), x(t))}{\partial \theta^{n+1}} \neq 0, \quad t \in [0, \tau(x_0)].$$

We next claim that $\dot{\theta}(x(t)) = -1$ almost everywhere on the segment $[0, \tau(x_0)]$. To do this, we argue by contradiction. Since from (2.10) it follows that $\dot{\theta}(x(t)) = -1$ when $x(t) \in \mathbb{M}_1$, our assumption means that $\mu\mathbb{Q} > 0$, where $\mathbb{Q} = \{t \in [0, \tau(x_0)] : x(t) \in \mathbb{M}_2, \dot{\theta}(x(t)) \neq -1\}$. Thus there exists a closed set $\mathbb{Q}^1 \subset \mathbb{Q}$, $\mu\mathbb{Q}^1 > 0$, which can be presented in the form of the union $\mathbb{Q}^1 = \mathbb{Q}_1 \cup \mathbb{Q}_2$, where \mathbb{Q}_1 is perfect and \mathbb{Q}_2 is at most countable.

LEMMA 2.3. *For every point (θ, x) , we have that*

$$(2.12) \quad \begin{aligned} & \Phi_{\theta}^{(m)}(\theta, x) - ([\Phi_x(\theta, x)]_{\theta}^{(m-1)}, Ax - BU^{-1}B^*N^{-1}(\theta)x) \\ &= \sum_{j=1}^{m-1} c_j^m (U^{-1/2}B^*[\Phi_x(\theta, x)]_{\theta}^{(m-j-1)}, U^{-1/2}[\Phi_x(\theta, x)]_{\theta}^{(j-1)}), \end{aligned}$$

where $c_j^m < 0$ ($j = 1, \dots, m-1$, $m = 2, 3, \dots$) is true.

Proof. Note first that, because of (2.5),

$$(2.13) \quad 2(Ax - BU^{-1}B^*N^{-1}(\theta)x)_{\theta} = -BU^{-1}B^*\Phi_x.$$

Then, from (2.4), we obtain that

$$\Phi_{\theta\theta} = (\Phi_{x\theta}, Ax - BU^{-1}B^*N^{-1}(\theta)x) - \frac{1}{2}(\Phi_x, BU^{-1}B^*\Phi_x),$$

which proves the validity of (2.12) in the case where $m = 2$. Let $m = k$ and assume that

$$\begin{aligned} & \Phi_{\theta}^{(k)} - ([\Phi_x]_{\theta}^{(k-1)}, Ax - BU^{-1}B^*N^{-1}(\theta)x) \\ &= \sum_{j=1}^{k-1} c_j^k (U^{-1/2}B^*[\Phi_x]_{\theta}^{(k-j-1)}, U^{-1/2}B^*[\Phi_x]_{\theta}^{(j-1)}), \end{aligned}$$

where $c_j^k < 0$, $j = 1, \dots, k-1$. Differentiating this relationship on θ and using (2.13), we obtain that

$$\begin{aligned} & \Phi_{\theta}^{(k+1)} - ([\Phi_x]_{\theta}^{(k)}Ax - BU^{-1}B^*N^{-1}(\theta)x) \\ &= \sum_{j=1}^k c_j^{k+1} (U^{-1/2}B^*[\Phi_x]_{\theta}^{(k-j)}, U^{-1/2}B^*[\Phi_x]_{\theta}^{(j-1)}), \end{aligned}$$

where $c_1^{k+1} = c_1^k - \frac{1}{2} < 0$, $c_j^{k+1} = c_{j-1}^k + c_j^k < 0$, $j = 2, \dots, k-1$, and $c_k^{k+1} = c_{k-1}^k < 0$. This proves the lemma by induction on m .

We return to the proof of the uniqueness of the solution of (2.1) when $x_0 \in \mathbb{M}_2$. For every $t \in \mathbb{Q}_1$, the definition of \mathbb{M}_2 yields

$$(2.14) \quad \Phi_{\theta}(\theta(x(t)), x(t)) = 0,$$

and, moreover,

$$(2.15) \quad \Phi_{\theta\theta}(\theta(x(t)), x(t)) = 0,$$

because $\theta(x(t))$ is the point where the function $\psi(\theta, x(t))$ achieves its global minimum. By differentiating identity (2.14) with respect to t at the points of the perfect set \mathbb{Q}_1 , we see that, for $t \in \mathbb{Q}_1$,

$$(\Phi_{\theta x}(\theta(x(t)), x(t)), Ax(t)BU^{-1}B^*N^{-1}(\theta(x(t)))x(t)) + \Phi_{\theta\theta}(\theta(x(t)), x(t))\dot{\theta} = 0.$$

Considering (2.15), we obtain that

$$(2.16) \quad (\Phi_{\theta x}(\theta(x(t)), x(t)), Ax(t) - BU^{-1}B^*N^{-1}(\theta(x(t)))x(t)) = 0, \quad t \in \mathbb{Q}_1.$$

Therefore Lemma 2.3 with $m = 2$, and (2.15), (2.16) imply that

$$B^*\Phi_x(\theta(x(t)), x(t)) = 0, \quad t \in \mathbb{Q}_1.$$

The next claim is that, for $n = 2k$, the inductive assumptions

$$(2.17) \quad \Phi_{\theta}(\theta(x(t)), x(t)) = \Phi_{\theta\theta}(\theta(x(t)), x(t)) = \dots = \Phi_{\theta}^{(n)}(\theta(x(t)), x(t)) = 0$$

and

$$(2.18) \quad B^*[\Phi_x(\theta(x(t)), x(t))]_{\theta}^{(i)} = 0, \quad i = 0, \dots, k-1, \quad t \in \mathbb{Q}_1$$

imply that

$$\Phi_{\theta}^{(n+1)}(\theta(x(t)), x(t)) = \Phi_{\theta}^{(n+2)}(\theta(x(t)), x(t)) = 0$$

and

$$B^*[\Phi_x(\theta(x(t)), x(t))]_{\theta}^{(k)} = 0, \quad t \in \mathbb{Q}_1.$$

To see this, use Lemma 2.3 with $m = n + 1$ and apply (2.18) to obtain that

$$\begin{aligned} & -([\Phi_x(\theta(x(t)), x(t))]_{\theta}^{(n)}, Ax(t) - BU^{-1}B^*N^{-1}(\theta(x(t)))x(t)) \\ & + \Phi_{\theta}^{(n+1)}(\theta(x(t)), x(t)) = 0, \quad t \in \mathbb{Q}_1. \end{aligned}$$

On the other hand, by differentiating (2.17) with respect to t , $t \in \mathbb{Q}_1$, we have that

$$\begin{aligned} & \Phi_{\theta}^{(n+1)}(\theta(x(t)), x(t))\dot{\theta}(x(t)) \\ & + ([\Phi_x(\theta(x(t)), x(t))]_{\theta}^{(n)}Ax(t) - BU^{-1}B^*N^{-1}(\theta(x(t)))x(t)) = 0. \end{aligned}$$

Adding the previous two equalities and considering that $\dot{\theta}(x(t)) \neq -1$, $t \in \mathbb{Q}_1$, we infer that

$$(2.19) \quad \Phi_{\theta}^{(n+1)}(\theta(x(t)), x(t)) = 0, \quad t \in \mathbb{Q}_1.$$

By (2.17), (2.19), and the fact that $\psi(\theta, x(t))$ has global minimum at the point $\theta(x(t))$, we also infer that

$$(2.20) \quad \Phi_{\theta}^{(n+2)}(\theta(x(t)), x(t)) = 0, \quad t \in \mathbb{Q}_1.$$

Furthermore, Lemma 2.3, with $m = n + 2$ and (2.18)–(2.20), yields

$$B^*[\Phi_x(\theta(x(t)), x(t))]_{\theta}^{(k)} = 0, \quad t \in \mathbb{Q}_1.$$

Thus we have proved by induction that, for every integer $n > 0$,

$$\Phi_{\theta}^{(n)}(\theta(x(t)), x(t)) = 0, \quad t \in \mathbb{Q}_1.$$

However, this clearly contradicts (2.11).

It follows that almost everywhere on the segment $[0, \tau(x_0)]$ we must have that $\dot{\theta}(x(t)) = -1$, and, because of absolute continuity, $\dot{\theta}(x(t)) = -1$ for every $t \in [0, \tau(x_0)]$. Therefore $x(t)$ is the solution of system (2.2) when $f(t) = -1$, which proves uniqueness for every $x_0 \in \mathbb{M}_2$, and completes the proof of Theorem 2.1.

3. Optimality of positional control. In this section, we will verify that the control $u(x)$, given by formula (1.6), also has properties (ii) and (iii) stated in the Introduction.

THEOREM 3.1. *Consider the optimal control problem (0.1)–(0.3). Let $\theta(x)$ be the controllability function and suppose that the control $u(x)$ is given by (1.6). Then, for every $x_0 \in \mathbb{R}^n$, we have that $\lim_{t \rightarrow \theta(x_0)} x(t) \rightarrow 0$, where $x(t)$ is the solution of the closed-loop system (0.4), and functional (0.3) achieves its minimum at this control $u(x(t))$.*

Proof. The fact that $\lim_{t \rightarrow \theta(x_0)} x(t) \rightarrow 0$ follows from

$$\lim_{t \rightarrow \theta(x_0)} \theta(x) = \lim_{t \rightarrow \theta(x_0)} (\theta(x_0) - t) = 0.$$

Consider the value of the cost functional (0.3) at the control $u(x)$ and the solution $x(t)$ corresponding to it. Using (2.5) and (2.6), we have that

$$\begin{aligned}
 J(u(x)) &= \lim_{\varepsilon \rightarrow +0} \int_0^{\theta(x_0) - \varepsilon} [(Uu(x(t)), u(x(t))) + (Wx(t), x(t))] dt + k^2 \theta(x_0) \\
 &= \lim_{\varepsilon \rightarrow +0} \int_0^{\theta(x_0) - \varepsilon} -\dot{\psi}(\theta(x(t)), x(t)) dt \\
 &= -\lim_{\varepsilon \rightarrow +0} \psi(\theta(x(\theta(x_0) - \varepsilon)), x(\theta(x_0) - \varepsilon)) + \psi(\theta(x_0), x_0) \\
 &= \psi(\theta(x_0), x_0) \\
 &= k^2 \theta(x_0) + (N^{-1}(\theta(x_0))x_0, x_0) \\
 &= \min_{\theta > 0} [k^2 \theta + (N^{-1}(\theta)x_0, x_0)].
 \end{aligned}$$

This completes the proof.

Example 3.1. Consider the optimal control problem

$$\dot{x} = u, \quad x(0) = x_0, \quad x(T) = 0, \quad J(u) = k^2 T + \int_0^T (x^2 + u^2) dt.$$

In this case, $N(t)$ is found as the solution of the Riccati equation

$$\dot{N} = 1 - N^2, \quad N(0) = 0,$$

and a simple computation yields

$$N(t) = \frac{e^{2t} - 1}{e^{2t} + 1}.$$

The functions ψ and Φ are given by

$$\psi(\theta, x) = k^2 \theta + \frac{e^{2\theta} + 1}{e^{2\theta} - 1} x^2, \quad \Phi(\theta, x) = k^2 - \frac{4 e^{2\theta}}{(e^{2\theta} - 1)^2} x^2.$$

The controllability function $\theta(x)$ is the solution of the equation

$$\frac{4 e^{2\theta} x^2}{(e^{2\theta} - 1) x^2} = k^2.$$

Solving this equation for θ , we obtain that

$$\theta = \frac{1}{2} \ln \frac{2x^2 + k^2 + 2\sqrt{x^4 + x^2 k^2}}{k^2}$$

(note that the other root satisfies $(2x^2 + k^2 - 2\sqrt{x^4 + x^2 k^2})/k^2 < 1$, and so discarded). Thus the optimal control is given by the formula

$$u(x) = -\frac{e^{2\theta(x)} + 1}{e^{2\theta(x)} - 1} x = -\frac{x^2 + k^2 + \sqrt{x^4 + x^2 k^2}}{x^2 + \sqrt{x^4 + k^2 x^2}} x.$$

4. Solution of the problem with the mixed cost function in the case where $W = 0$. In the case considered in the previous two sections, the solution of the differential Riccati equation (1.2), as well as the solution of the differential equation (0.4) in the sense of

Definition 2.1, are not presented in the explicit form. If, however, we restrict our attention to the case where $W = 0$, then explicit forms of the solutions can be obtained. The Riccati equation (1.2) in this case has the form

$$(4.1) \quad \dot{N} = -AN - NA^* + BU^{-1}B^*,$$

and its solution for the initial condition $N(0) = 0$ is given by

$$(4.2) \quad N(t) = \int_0^t e^{-A\tau} BU^{-1}B^* e^{-A^*\tau} d\tau.$$

Hence it follows that this matrix is positive definite for $t > 0$ and satisfies

$$(4.3) \quad \begin{aligned} AN(t) + N(t)A^* &= A \int_0^t e^{-A\tau} BU^{-1}B^* e^{-A^*\tau} d\tau + \int_0^t e^{-A\tau} BU^{-1}B^* e^{-A^*\tau} d\tau A^* \\ &= - \int_0^t dI(t) = BU^{-1}B^* - I(t), \end{aligned}$$

where

$$(4.4) \quad I(t) = e^{-At} BU^{-1}B^* e^{-A^*t}.$$

The inverse matrix is readily seen to satisfy the equation

$$(4.5) \quad N^{-1}(t)A + A^*N^{-1}(t) = N^{-1}(t)BU^{-1}B^*N^{-1}(t) - N^{-1}(t)I(t)N^{-1}(t),$$

and the functions ψ and Φ take the form

$$(4.6) \quad \Phi(\theta, x) = k^2 - (I(\theta)N^{-1}(\theta)x, N^{-1}(\theta)x),$$

$$(4.7) \quad \psi(\theta, x) = k^2\theta + (N^{-1}(\theta)x, x), \quad k \neq 0.$$

Using these expressions to help in defining the controllability function $\theta(x)$, and using the control $u(x) = -U^{-1}B^*N^{-1}(\theta(x))x$, we can easily obtain the explicit form of the solution of (2.1) in the sense of Definition 2.1.

In fact, considering that on the trajectories of (2.1) the function $\theta(x)$ satisfies the equation $\dot{\theta}(x(t)) = -1$, and substituting $y(t) = N^{-1}(\theta(x(t)))x(t)$, we have that

$$\begin{aligned} \dot{y} &= [N^{-1}(\theta(x(t)))A + N^{-1}(\theta(x(t)))I(\theta(x(t)))N^{-1}(\theta(x(t))) \\ &\quad - N^{-1}(\theta(x(t)))BU^{-1}B^*N^{-1}(\theta(x(t)))]x(t). \end{aligned}$$

Because of identity (4.5) we obtain that $\dot{y}(t) = -A^*y(t)$. Hence

$$y(t) = e^{-A^*t} N^{-1}(\theta(x_0))x_0$$

and

$$x(t) = N(\theta(x(t))) e^{-A^*t} N^{-1}(\theta(x_0))x_0 = N(\theta(x_0) - t) e^{-A^*t} N^{-1}(\theta(x_0))x_0.$$

Thus $x(t) \rightarrow 0$ when $t \rightarrow \theta(x_0)$ because $N(\theta(x_0) - t) \rightarrow 0$ as $t \rightarrow \theta(x_0)$.

Example 4.1. Consider the optimal control problem

$$\begin{aligned} \dot{x}_1 &= x_2, & \dot{x}_2 &= u, \\ J(u) &= k^2 T + \int_0^T u^2 dt; \end{aligned}$$

that is, $\dot{x} = Ax + Bu$, where

$$\begin{aligned} x \in \mathbb{R}^2, \quad u \in \mathbb{R}^1, \quad A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \\ I(\theta) = e^{-A\theta} B B^* e^{-A^*\theta} = \begin{pmatrix} \theta^2 & -\theta \\ -\theta & 1 \end{pmatrix}, \\ N(\theta) = \int_0^\theta e^{-A t} B B^* e^{-A^* t} dt = \begin{pmatrix} \theta^3/3 & -\theta^2/2 \\ -\theta^2/2 & \theta \end{pmatrix}, \\ N^{-1}(\theta) = \frac{12}{\theta^4} \begin{pmatrix} \theta & \theta^2/2 \\ \theta^2/2 & \theta^3/3 \end{pmatrix}. \end{aligned}$$

A direct computation yields, for $\theta > 0$, that

$$\begin{aligned} \psi(\theta, x) &= (N^{-1}(\theta)x, x) + k^2\theta = \frac{12x_1^2}{\theta^3} + \frac{12x_1x_2}{\theta^2} + \frac{4x_2^2}{\theta} + k^2\theta, \quad k > 0, \\ \Phi(\theta, x) &= k^2 - (N^{-1}(\theta)J(\theta)N^{-1}(\theta)x, x) \\ &= k^2 - \frac{36x_1^2}{\theta^4} - \frac{24x_1x_2}{\theta^3} - \frac{4x_2^2}{\theta^2} \\ &= k^2 - \left(6\frac{x_1}{\theta^2} + 2\frac{x_2}{\theta}\right)^2, \quad k > 0. \end{aligned}$$

The function $\theta = \theta(x)$ is found as the minimal solution of the equation

$$(4.8) \quad k^2\theta = 36x_1^2 + 24x_1x_2\theta + 4x_2^2\theta^2,$$

at which the function

$$(4.9) \quad \psi(\theta, x) = \frac{12x_1^2}{\theta^3} + \frac{12x_1x_2}{\theta^2} + \frac{4x_2^2}{\theta} + k^2\theta$$

assumes into its global minimum.

Consider the parabola defined by the equation

$$(4.10) \quad x_1 = -\alpha_0 x_2 |x_2|,$$

where α_0 is the unique root of the equation

$$(4.11) \quad \begin{aligned} G(\alpha) &= k^2[b_1 - b_2] + 4\left[\frac{1}{b_1} - \frac{1}{b_2}\right] - 12\alpha\left[\frac{1}{b_1^2} - \frac{1}{b_2^2}\right] + 12\alpha^2\left[\frac{1}{b_1^3} - \frac{1}{b_2^3}\right] = 0, \\ b_1 &= \frac{1}{k}[\sqrt{1-6k\alpha} - 1], \quad b_2 = \frac{1}{k}[\sqrt{1-6k\alpha} + 1], \end{aligned}$$

chosen so that $0 < \alpha_0 < 1/6k$. To see that such a root exists, observe that if $\alpha \ll 1$, then $G(\alpha) \approx 4/9\alpha > 0$, while if $\alpha = 1/6k$, then $G(\alpha) = \frac{4}{3}[2\sqrt{2} - 3] < 0$. To show that the root α_0 is unique, we calculate

$$\begin{aligned} G'(\alpha) &= \frac{2\alpha - b_1}{b_1^3} - \frac{2\alpha - b_2}{b_2^3} \\ &= \frac{2\alpha}{k^2} [2(\sqrt{1+6k\alpha} - \sqrt{1-6k\alpha} + \sqrt{1-36k^2\alpha^2} - 3)] \\ &= \frac{2\alpha}{k^2} P(\alpha). \end{aligned}$$

It is evident that

$$P(0) = -3, \quad P\left(\frac{1}{6k}\right) = 2\sqrt{2} - 3 < 0,$$

and

$$P'(\alpha) = \frac{6k}{\sqrt{1+6k\alpha}\sqrt{1-6k\alpha}} [\sqrt{1+6k\alpha} + \sqrt{1-6k\alpha} - 6k\alpha] > 0$$

for every $\alpha \in (0, 1/6k)$. Therefore $P(\alpha) < 0$ for every $\alpha \in (0, 1/6k)$; that is, $G'(\alpha) < 0$, for every $\alpha \in (0, 1/6k)$, so (4.11) has a unique root in the interval $(0, 1/6k)$.

Let

$$\begin{aligned} S_1 &= \{x \in \mathbb{R}^2: x_1 = -\alpha_0 x_2 |x_2|, x_1 > 0\}, \\ S_2 &= \{x \in \mathbb{R}^2: x_1 = -\alpha_0 x_2 |x_2|, x_1 < 0\}, \\ S^+ &= \{x \in \mathbb{R}^2: x_1 > -\alpha_0 x_2 |x_2|\}, \quad S_- = \{x \in \mathbb{R}^2: x_1 < -\alpha_0 x_2 |x_2|\}, \\ (4.12) \quad \theta(x) &= \begin{cases} 1/k(x_2 + \sqrt{x_2^2 + 6kx_1}), & (x_1, x_2) \in S^+ \cup S_1, \\ 1/k(-x_2 + \sqrt{x_2^2 - 6kx_1}), & (x_1, x_2) \in S^- \cup S_2. \end{cases} \end{aligned}$$

The function $\theta(x)$ has a discontinuity at the point of the set $S_1 \cup S_2$ because

$$\begin{aligned} \frac{1}{k}(x_2 + \sqrt{x_2^2 + 6kx_1}) &> \frac{1}{k}(-x_2 + \sqrt{x_2^2 - 6kx_1}), \quad (x_1, x_2) \in S_2, \\ \frac{1}{k}(-x_2 + \sqrt{x_2^2 - 6kx_1}) &> \frac{1}{k}(x_2 + \sqrt{x_2^2 + 6kx_1}), \quad (x_1, x_2) \in S_1. \end{aligned}$$

It is evident that, in this case, the set

$$\mathbb{M}_2 = \{x \in \mathbb{R}^2 \setminus \{0\}: \Phi'_\theta(\theta(x), x) = 0\}$$

introduced in § 2 is empty. The control $u(x)$ transferring an arbitrary space point \mathbb{R}^2 to the origin is given explicitly by

$$(4.13) \quad u(x) = -B^* N^{-1}(\theta(x))x = -\frac{6}{\theta^2(x)} x_1 - \frac{4}{\theta(x)} x_2;$$

that is,

$$(4.14) \quad u(x) = \begin{cases} \frac{x_2^2 - x_2 \sqrt{x_2^2 + 6kx_1}}{3x_1} - k, & (x_1, x_2) \in S^+ \cup S_1, \\ \frac{x_2^2 + x_2 \sqrt{x_2^2 - 6kx_1}}{3x_1} + k, & (x_1, x_2) \in S^- \cup S_2. \end{cases}$$

The solution of the system

$$(4.15) \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u(x),$$

in the sense of Definition 2.1, where $u(x)$ is given by (4.14), is given by formulas

$$\begin{aligned} x_1(t) &= c_1(\theta(x_0) - t)^2 + c_2(\theta(x_0) - t)^3, \\ x_2(t) &= -2c_1(\theta(x_0) - t) - 3c_2(\theta(x_0) - t)^2, \end{aligned}$$

where c_1, c_2 are found from initial conditions.

We note that if the point $x_0 = (x_{10}, x_{20}) \in S_1 \cup S_2$, then there exist two trajectories satisfying system (4.15) for every t and leading to 0; namely,

$$(4.16) \quad x_1(t) = c_1(a_1 - t)^2 + c_2(a_1 - t)^3, \quad x_2(t) = -2c_1(a_1 - t) - 3c_2(a_1 - t)^2,$$

where

$$a_1 = \frac{1}{k} (x_{20} + \sqrt{x_{20}^2 + 6kx_{10}}), \quad c_1 = \frac{a_1 x_{20} - 3\alpha_0 x_{20} |x_{20}|}{a_1^2},$$

$$c_2 = \frac{-a_1 x_{20} + 2\alpha_0 x_{20} |x_{20}|}{a_1^3},$$

and

$$(4.17) \quad x_1(t) = \tilde{c}_1(a_2 - t)^2 + \tilde{c}_2(a_2 - t)^3, \quad x_2(t) = -2\tilde{c}_1(a_2 - t) - 3\tilde{c}_2(a_2 - t)^2,$$

where

$$a_2 = \frac{1}{k} (-x_{20} + \sqrt{x_{20}^2 - 6kx_{10}}), \quad \tilde{c}_1 = \frac{a_2 x_{20} - 3\alpha_0 x_{20} |x_{20}|}{a_2^2},$$

$$\tilde{c}_2 = \frac{-a_2 x_{20} + 2\alpha_0 x_{20} |x_{20}|}{a_2^3}.$$

Furthermore, it is easy to see that when $x_0 \in S_1$, trajectory (4.16) satisfies Definition 2.1. On trajectory (4.17), the function $\theta(x(t))$ is not absolutely discontinuous, although the equality $\dot{\theta}(x(t)) = -1$ is satisfied for every t . For $x_0 \in S_2$, the only trajectory that satisfies Definition 2.1 is trajectory (4.17). The constructed synthesis gives the optimal solution of the problem

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u,$$

$$(4.18) \quad J(u) = k^2 T + \int_0^T u^2 dt \rightarrow \inf.$$

The optimal value of the cost function is

$$J_{x_0}^* = \frac{12x_{10}^2}{\theta^3(x_0)} + \frac{12x_{10}x_{20}}{\theta^2(x_0)} + \frac{4x_{20}^2}{\theta(x_0)} + k^2\theta(x_0),$$

where $\theta(x)$ is given by (4.12). From (4.11), it follows that, for $x_0 \in S_1 \cup S_2$, both trajectories (4.16) and (4.17) are solutions of problem (4.18).

We conclude this section by noting that the forms of (4.8) for the definition of the controllability function and of control (4.13) are fully analogous with the forms of the equation for the function θ and for the control $u(x)$ of a corresponding example from the work [4], while solving the problem of finding possible nonoptimal positional control.

5. Finding the positional control in the optimal stabilizability problem. In this section, the connection between the solution of the problem with the mixed cost function (0.1)–(0.3) and the optimal stabilizability problem is investigated.

THEOREM 5.1. *Let $W > 0$. Then, for every $x_0 \in \mathbb{R}^n$, the optimal control $u_k(x)$ for problem (0.1)–(0.3) tends to the optimal control $u_0(x)$ for the optimal stabilizability problem when $k \rightarrow 0$.*

Proof. Let $N(t)$ be the solution of the Cauchy problem (1.2) and let N_1 be a positive definite matrix solution of the algebraic Riccati equation

$$(5.1) \quad AN_1 + N_1A^* - BU^{-1}B^* + N_1WN_1 = 0.$$

The matrix G given by

$$(5.2) \quad G = N_1 - N(t)$$

is the solution of the Cauchy problem

$$(5.3) \quad \dot{G} = -(A + N_1 W)G - G(A^* + WN_1) + G W G, \quad G(0) = N_1.$$

In addition to (5.3), we consider the Cauchy problem

$$(5.4) \quad \dot{P} = P(A + N_1 W) + (A^* + WN_1)P - W, \quad P(0) = N_1^{-1}.$$

As is well known, the solution of this problem is

$$(5.5) \quad P(t) = e^{(A^* + WN_1)t} N_1^{-1} e^{(A + N_1 W)t} - \int_0^t e^{(A^* + WN_1)\tau} W e^{(A + N_1 W)\tau} d\tau.$$

It will be convenient to introduce the function

$$(5.6) \quad \begin{aligned} \dot{f}_x(t) = (P(t)x, x) &= (N_1^{-1} e^{(A + N_1 W)t} x, e^{(A + N_1 W)t} x) \\ &- \int_0^t (W e^{(A + N_1 W)\tau} x, e^{(A + N_1 W)\tau} x) d\tau \end{aligned}$$

for $x \neq 0$ and $t \geq 0$.

We structure the remainder of the proof as a sequence of lemmas.

LEMMA 5.1. *For every fixed $x \neq 0$, we have that $\dot{f}_x(t) \geq 0$, for every $t \geq 0$, and, for every $t_1 > 0$, $\dot{f}_x(t)$ does not vanish identically on the interval $[0, t_1]$.*

Proof. We first calculate

$$\dot{f}_x(t) = (N_1^{-1}(A + N_1 W)y, y) + (N_1^{-1}y, (A + N_1 W)y) - (Wy, y),$$

where $y = e^{(A + N_1 W)t} x$. Then, because of (5.1),

$$(5.7) \quad \begin{aligned} \dot{f}_x(t) &= ([N_1^{-1}A + W + A^*N_1^{-1} + W - W]y, y) \\ &= (N_1^{-1}BU^{-1}B^*N_1^{-1}y, y) \\ &= \|U^{-1/2}B^*N_1^{-1}e^{(A + N_1 W)t}x\|^2 \\ &\geq 0. \end{aligned}$$

Suppose that $x \neq 0$ and $t_1 > 0$ are such that

$$(5.8) \quad B^*N_1^{-1}e^{(A + N_1 W)t}x = 0 \quad \text{for every } t \in [0, t_1].$$

Using (5.1), repeatedly differentiating (5.8), and setting $t = 0$, we obtain that

$$\begin{aligned} B^*(N_1^{-1}A + W)x &= B^*(A^*N_1^{-1} - N_1^{-1}BU^{-1}B^*N_1^{-1})x \\ &= B^*(A^* - N_1^{-1}BU^{-1}B^*)N_1^{-1}x = 0, \\ B^*(A^* - N_1^{-1}BU^{-1}B^*)^2N_1^{-1}x &= 0, \\ &\vdots \\ B^*(A^* - N_1^{-1}BU^{-1}B^*)^{m-1}N_1^{-1}x &= 0. \end{aligned}$$

Hence, because of complete controllability of the pair $((A - BU^{-1}B^*)N_1^{-1}, B)$, it follows that $N_1^{-1}x = 0$, $x \neq 0$, which contradicts to positive definiteness of N_1 . This completes the proof of Lemma 5.1.

COROLLARY 5.1. *At every fixed $x \neq 0$, the function $(P(t)x, x)$ is strictly increasing for $t > 0$; i.e., $t_2 > t_1$ implies that $(P(t_2)x, x) > (P(t_1)x, x)$.*

COROLLARY 5.2. *It holds that $N(t) \rightarrow N_1$ as $t \rightarrow +\infty$.*

To prove the corollaries, note that, because of the continuity of $\dot{f}_x(t)$ for each fixed x , there exists for every $T > 0$ an interval $[\alpha, \beta] \subset [0, T]$ on which $\dot{f}_x(t) > 0$. Consequently,

$$(5.9) \quad \int_0^T \dot{f}_x(t) dt > 0 \quad \text{for every } x \neq 0, T > 0.$$

If we designate

$$(5.10) \quad M(t) = \int_0^t e^{(A^* + WN_1)\tau} BU^{-1} B^* e^{(A + N_1 W)\tau} d\tau,$$

then, from (5.9), it follows that $M(t)$ is positive definite for $t > 0$. Choose a constant $c > 0$ such that, for every $x \in \mathbb{R}^n$,

$$(5.11) \quad (M(1)x, x) = \left(\int_0^1 e^{(A^* + WN_1)\tau} BU^{-1} B^* e^{(A + N_1 W)\tau} d\tau x, x \right) \geq c \|x\|^2.$$

We claim that $(M(t)x, x) \rightarrow +\infty$ as $t \rightarrow +\infty$. In fact,

$$(5.12) \quad \begin{aligned} (M(t)x, x) &= \sum_{k=0}^{[t]} \left(\int_0^1 e^{(A^* + WN_1)\tau} BU^{-1} B^* e^{(A + N_1 W)\tau} d\tau x, x \right) \\ &\quad + \left(\int_{[t]}^t e^{(A^* + WN_1)\tau} BU^{-1} B^* e^{(A + N_1 W)\tau} d\tau x, x \right) \\ &= \sum_{k=0}^{[t]} \left(\int_0^1 e^{(A^* + WN_1)\sigma} BU^{-1} B^* e^{(A + N_1 W)\sigma} d\sigma e^{(A + N_1 W)k} x, e^{(A + N_1 W)k} x \right) \\ &\quad + \left(\int_{[t]}^t e^{(A^* + WN_1)\tau} BU^{-1} B^* e^{(A + N_1 W)\tau} d\tau x, x \right). \end{aligned}$$

It is known [10] that $A - BU^{-1} B^* N_1^{-1}$ is a stable matrix; therefore, from (5.1), it follows that $-A^* - WN_1$ is also stable. Thus there exists a constant $c_1 > 0$ such that $\|e^{(A + N_1 W)k} x\| \leq c_1 \|x\|$ for every $k > 0$ and for every x . From (5.11) and (5.12), it follows that $(M(t)x, x) \geq [t]cc_1^2 \|x\|^2$.

Considering relationship (5.6), we have, for $t > 1$, that

$$(P(t)x, x) = \int_0^t \dot{f}_x(\tau) d\tau = (M(t)x, x) + (N_1^{-1}x, x) \geq [t]cc_1^2 \|x\|^2.$$

Thus, for the inverse matrix $G(t) = P(t)^{-1}$, we have that

$$(G(t)x, x) \leq \|x\|^2 / ([t]cc_1^2),$$

which immediately yields that $(G(t)x, x) \rightarrow 0$ as $t \rightarrow \infty$. The conclusion of the corollary follows from relationship (5.2).

LEMMA 5.2. *Let $Q(t)$ be a positive definite matrix for $t > 0$ and suppose that, for each $x \neq 0$, the function $(Q(t)x, x)$ is strictly increasing for $t > 0$. Then, for each $x \neq 0$, the function $(Q^{-1}(t)x, x)$ is strictly decreasing for $t > 0$.*

Proof. A direct computation yields that

$$\frac{d}{dt} (Q^{-1}(t)x, x) = -(Q^{-1} \dot{Q} Q^{-1} x, x) = -(\dot{Q}(t) Q^{-1}(t)x, Q^{-1}(t)x) = -(\dot{Q}(t)y, y) \leq 0,$$

where $y = Q^{-1}(t)x$; thus the function $(Q^{-1}(t)x, x)$ is nonincreasing.

To show that this function is strictly decreasing, suppose, to the contrary, that there exist t_1 and t_2 , $t_1 > t_2$ such that $(Q^{-1}(t_2)x, x) = (Q^{-1}(t_1)x, x)$ for some $x \neq 0$; i.e., $([Q^{-1}(t_2) - Q^{-1}(t_1)]x, x) = 0$. Since $Q^{-1}(t_2) - Q^{-1}(t_1)$ is negative semidefinite, it follows that $Q^{-1}(t_2)x = Q^{-1}(t_1)x$. Designating $Q^{-1}(t_1)x = y$, we obtain that $x = Q(t_2)y = Q(t_1)y$ and $(Q(t_2)y, y) = (Q(t_1)y, y)$, which contradicts the lemma's supposition. This completes the proof.

COROLLARY 5.3. *For every fixed $x \neq 0$, the function $(N^{-1}(t)x, x)$ is strictly decreasing for $t > 0$.*

In fact, from Corollary 5.1, we have that $(P(t)x, x)$ is strictly increasing. Since $P(t)$ is positive definite for $t = 0$, it follows that $P(t)$ is positive definite for $t > 0$. Thus the inverse matrix $P^{-1}(t)$ is positive definite, and, according to (5.2) and (5.3), coincides with $G(t)$. Considering relationship (5.2) and applying Lemma 5.2, we twice obtain the corollary.

LEMMA 5.3. *For x fixed, the controllability function $\theta_k(x)$ tends to $+\infty$ as $k \rightarrow 0$.*

Proof. For $T > 0$ and $\theta \in [0, T]$, we have that

$$\psi(\theta, x) = k^2\theta + (N^{-1}(\theta)x, x) \geq (N^{-1}(\theta)x, x) \geq (N^{-1}(T)x, x).$$

In accordance with Corollary 5.3, $(N^{-1}(T)x, x) > (N^{-1}(2T)x, x)$, so there exists $k > 0$ such that

$$k^2T + (N^{-1}(2T)x, x) < (N^{-1}(T)x, x) \leq \psi(\theta, x) \quad \text{for every } \theta \in [0, T].$$

Thus $\psi(\theta, x)$ cannot attain its global minimum on the interval $[0, T]$. The statement of lemma follows because $T > 0$ is arbitrary.

Theorem 5.1 follows from Lemma 5.3 and Corollary 5.2, since

$$J(u) \rightarrow \int_0^T [(Wx, x) + (Uu, u)] dt \quad \text{as } k \rightarrow 0,$$

and, because of (1.6), the control $u(x) \rightarrow -U^{-1}B^*N_1x$ as $k \rightarrow 0$.

If we drop the assumption that $W > 0$, then the matrix equation (5.1) may have no solution. Nevertheless, the limit of the matrix $N^{-1}(\theta)$ can exist, and, moreover, the control $u = -U^{-1}B^*Mx$, where $M = \lim_{\theta \rightarrow \infty} N^{-1}(\theta)$, can be optimal for the linear quadratic problem; that is, the form of Theorem 5.1 can occur when positive definiteness of the matrix W is no longer assumed.

We will consider the case when $W = 0$ and will obtain the exact answer to the question about the validity limits of Theorem 5.1 in terms of the matrix spectrum. With this purpose we consider the matrix

$$N(\theta) = \int_0^\theta e^{-A\tau} B U^{-1} B^* e^{-A^*\tau} d\tau$$

as the solution for the differential equation (4.1), satisfying identity (4.5).

We claim that $\lim_{\theta \rightarrow \infty} N^{-1}(\theta) = M$ exists and that this limit is the solution of the equation

$$(5.13) \quad MA + A^*M = MBU^{-1}B^*M.$$

LEMMA 5.4. *The function $N^{-1}(\theta)$ has a limit $M \in \mathbb{R}^{n^2}$ as $\theta \rightarrow \infty$, and M satisfies (5.13).*

Proof. Note that the function $(N^{-1}(\theta)x, x)$ is decreasing in θ for every fixed $x \in \mathbb{R}^n$ because, due to (4.4), its derivative $-(N^{-1}(\theta)I(\theta)N^{-1}(\theta)x, x)$ is nonpositive. Since

the function $(N^{-1}(\theta)x, x)$ is also bounded from below, the matrix $N^{-1}(\theta)$ has a limit M when $\theta \rightarrow \infty$.

Let us further consider identity (4.5). Letting $\theta \rightarrow \infty$ in (4.5), we know that the matrix $N^{-1}(\theta)I(\theta)N^{-1}(\theta)$ has a limit, and therefore the function $-(N^{-1}(\theta)I(\theta)N^{-1}(\theta)x, x)$ also has some limit. Since the latter is the derivative of the function $(N^{-1}(\theta)x, x)$, which has a finite limit, it follows that

$$\lim_{\theta \rightarrow \infty} (N^{-1}(\theta)I(\theta)N^{-1}(\theta)x, x) = 0$$

for every $x \in \mathbb{R}^n$; hence

$$\lim_{\theta \rightarrow \infty} (N^{-1}(\theta)I(\theta)N^{-1}(\theta) = 0.$$

Therefore the matrix M satisfies (5.13), and Lemma 5.4 is proved.

Let $\tilde{A} = A - BU^{-1}B^*M$. The root subspaces of an $n \times n$ matrix C corresponding to the eigenvalues with the positive, negative, and zero real parts are denoted by $L_+(C)$, $L_-(C)$, and $L_0(C)$, respectively.

LEMMA 5.5. *It holds that $L_+(\tilde{A}) \cap \text{Ker } M = \{0\}$; $L_+(A) \cap \text{Ker } M = \{0\}$.*

Proof. We argue by contradiction. Let x be an eigenvector of the matrix A , or \tilde{A} , corresponding to an eigenvalue λ with $\text{Re } \lambda > 0$, and suppose that $Mx = 0$. Then $\tilde{A}x = Ax = \lambda x$. Choose a control $u(t)$, $t \in [0, T]$ that transfers the point $x \in \mathbb{R}^n$ to the origin during the time interval $[0, T]$. We have that $-x = \int_0^T e^{-A^*t} Bu(t) dt$. Let y be a root vector of the matrix A^* corresponding to the eigenvalue λ such that $(x, y) \neq 0$. Then there exists a certain vector polynomial $q(t)$ such that

$$\begin{aligned} |(x, y)| &= \left| \int_0^T (e^{-A^*t} Bu(t), y) dt \right| = \left| \int_0^T (u(t), B^* e^{-\bar{\lambda}t} q(t)) dt \right| \\ (5.14) \quad &\leq \left(\int_0^T (Uu(t), u(t)) dt \right)^{1/2} \left(\int_0^\infty \|B^* e^{-\bar{\lambda}t} q(t)\|^2 dt \right)^{1/2} \\ &\leq c \left(\int_0^T (Uu(t), u(t)) dt \right)^{1/2}, \end{aligned}$$

where $c = \left(\int_0^\infty \|B^* e^{-\bar{\lambda}t} q(t)\|^2 dt \right)^{1/2}$.

Specifically, let us choose $u(t)$ in the form of $u(t) = -U^{-1}B^* e^{-A^*t} N^{-1}(T)x$. It is clear that this control transfers the point x to the origin. Then, from (5.14), we have that

$$\begin{aligned} \frac{1}{c} |(x, y)| &\leq \left(\int_0^T (Uu(t), u(t)) dt \right)^{1/2} \\ &= \left(\int_0^T (B^* e^{-A^*t} N^{-1}(T)x, U^{-1}B^* e^{-A^*t} N^{-1}(T)x) dt \right)^{1/2} \\ &= (N(T)N^{-1}(T)x, N^{-1}(T)x)^{1/2} = (N^{-1}(T)x, x)^{1/2}; \end{aligned}$$

hence

$$(5.15) \quad (Mx, x) \geq \frac{1}{c^2} |(x, y)|^2 > 0.$$

Inequality (5.15) contradicts the assumption that $Mx = 0$, and the lemma is proved.

LEMMA 5.6. *It is true that $L_0(\tilde{A}) + L_+(\tilde{A}) \subset \text{Ker } M$.*

Proof. Let x be an arbitrary eigenvector of the matrix \tilde{A} corresponding to an eigenvalue λ with $\text{Re } \lambda \geq 0$. We will show that $Mx = 0$.

We argue by contradiction. Let $Mx \neq 0$. Then, from relationship (5.13), it follows that

$$(5.16) \quad A^*Mx = -\lambda Mx.$$

Taking the inner product of the equality $Ax - BU^{-1}B^*Mx = \lambda x$ with Mx and using (5.16), we have that

$$(5.17) \quad (Ax, Mx) - (BU^{-1}B^*Mx, Mx) = -\bar{\lambda}(Mx, x) = -\|U^{-1/2}B^*Mx\|^2.$$

Because of (5.17), $B^*Mx = 0$. Since $Mx \neq 0$ is an eigenvector of the matrix A^* , it follows that $B^*(A^*)^i Mx = 0$, $i = 1, 2, 3, \dots$, which contradicts the complete controllability of the system. Therefore $Mx = 0$.

The remainder of the proof will be performed by induction. Suppose it has been proved that, for all root vectors x of height r of the matrix \tilde{A} located in $L_0(\tilde{A}) + L_+(\tilde{A})$, $Mx = 0$ is fulfilled. We will prove that $My = 0$, where $y \in L_0(\tilde{A}) + L_+(\tilde{A})$ is an arbitrary root vector of height $r+1$. Again, we argue by contradiction. Let $My \neq 0$. Then, from relationship (5.13) and from the assumption of induction, we have that

$$(5.18) \quad -(A^* + \lambda I)My = M(\tilde{A} - \lambda I)y = 0;$$

that is, My is an eigenvector of A^* corresponding to the value eigenvalue λ .

As shown before, from the equality $Ay - BU^{-1}B^*My = \lambda y + x$, where x is a root vector of height r , and using (5.18), we obtain that

$$\begin{aligned} (Ay, My) - (BU^{-1}B^*My, My) &= -\bar{\lambda}(My, y) - \|U^{-1/2}B^*My\|^2 \\ &= \lambda(My, y) + (x, My) = \lambda(My, y), \end{aligned}$$

for $(x, My) = (Mx, y) = 0$ because of the induction assumption. Hence

$$0 \leq (\bar{\lambda} + \lambda)(My, y) = \|U^{-1/2}B^*My\|^2,$$

which contradicts the complete controllability of the system. This proves the lemma.

LEMMA 5.7. *It holds that $\text{Im } M \subset L_+(A^*)$.*

Proof. Let y_1, \dots, y_n be a basis of \mathbb{R}^n that consists of the proper and root vectors of the matrix \tilde{A} . Then

$$\text{Im } M = \text{Span}(My_1, \dots, My_n).$$

Suppose, for some i , $1 \leq i \leq n$, that $My_i \neq 0$ and let λ_i be the eigenvalue corresponding to the root vector y_i . Because of Lemma 5.6, $\text{Re } \lambda_i < 0$. We have, using relation (5.13), that

$$\begin{aligned} (A^* + \lambda I)My_i &= (A^* + \lambda I)My_i^0 = My_i^1, \\ &\vdots \\ (A^* + \lambda I)My_i^{r-1} &= 0, \end{aligned}$$

where r is the height of the root vector y_i . Then $(A^* + \lambda I)^r My_i = 0$, and therefore $My_i \in L_+(A^*)$.

Because of the arbitrariness of i , $1 \leq i \leq n$, the lemma is proved.

THEOREM 5.2. *The spectrum of the matrix $\tilde{A} = A - BU^{-1}B^*M$ is contained in the closed left half plane; that is, the state space \mathbb{R}^n can be represented in form of direct sum $\mathbb{R}^n = L_-(\tilde{A}) + L_0(\tilde{A})$, where $L_0(\tilde{A}) = L_0(A)$, $L_-(\tilde{A}) \supset L_-(A)$.*

Proof. From Lemmas 5.5 and 5.6, it directly follows that $L_+(\tilde{A}) = \{0\}$, so that $\mathbb{R}^n = L_-(\tilde{A}) + L_0(\tilde{A})$. From Lemma 5.6, it also follows that $L_0(\tilde{A}) \subset \text{Ker } M$, which leads to the inclusion

$$(5.19) \quad L_0(\tilde{A}) \subseteq L_0(A).$$

In fact, $\tilde{A}x = Ax$ for every $x \in L_0(\tilde{A})$, which yields that

$$\tilde{A}^2x = A^2x + BU^{-1}B^*MAx = A^2x + BU^{-1}B^*M\tilde{A}x = A^2x.$$

Similarly, $\tilde{A}^nx = A^nx$, for every $n = 3, 4, \dots$, so the given inclusion (5.19) is proved.

Furthermore, from Lemma 5.5, it follows that

$$\dim L_+(A) = \dim L_+(A^*) \leq \dim \text{Im } M.$$

However, from Lemma 5.7, we conclude that $\text{Im } M = L_+(A^*)$. Hence $\text{Ker } M = L_-(A) + L_0(A)$, and, as we saw before,

$$(5.20) \quad L_0(A) \subseteq L_0(\tilde{A}),$$

$$(5.21) \quad L_-(A) \subseteq L_-(\tilde{A}).$$

The assertion of the theorem follows from the inclusions (5.19)–(5.21).

THEOREM 5.3 (concerning the solution of the optimal stabilizability problem when $W = 0$). 1. *For any initial data $x_0 \in L_-(A)$, the control $u = u(x) = -U^{-1}B^*Mx$ gives the solution of the optimal stabilizability problem.*

2. *If $x_0 \notin L_-(A)$, an optimal solution does not exist.*

Proof. Let us denote $m(x) = \inf_{u(t) \in \Omega} J(u)$, where $J(u) = \int_0^\infty (Uu(t), u(t)) dt$ and Ω is the set of bounded measurable controls that transfer the initial state x_0 asymptotically to the origin as $t \rightarrow \infty$.

We claim that $m(x_0) = (Mx_0, x_0)$. To see this, let

$$(5.22) \quad u_k(t) = \begin{cases} -U^{-1}B^*e^{-A^*t}N^{-1}(k)x_0, & t \in [0, k], \\ 0, & t \in (k, \infty), \end{cases}$$

for $k = 1, 2, \dots$. Then $J(u_k) = (N^{-1}(k)x_0, x_0)$, $k = 1, 2, \dots$, and hence

$$(5.23) \quad m(x_0) \leq (Mx_0, x_0).$$

Suppose that $m(x_0) < (Mx_0, x_0)$. Then there exist $\varepsilon > 0$ and a control $\bar{u}(t)$ transferring the point x_0 asymptotically to the origin such that

$$(5.24) \quad m(x_0) \leq \int_0^\infty (U\bar{u}(t), \bar{u}(t)) dt < (Mx_0, x_0) - \varepsilon.$$

Let $\bar{x}(t)$ be the solution of the equation $\dot{\bar{x}}(t) = A\bar{x}(t) + B\bar{u}(t)$ with initial condition x_0 . Since $\bar{x}(t) \rightarrow 0$ as $t \rightarrow \infty$, then there exists $T > 0$ such that

$$(5.25) \quad (N^{-1}(1)\bar{x}(T), \bar{x}(T)) < \varepsilon.$$

Let the control $\bar{\bar{u}}(t)$ be defined on the segment $[0, T+1]$ by the formula

$$\bar{\bar{u}}(t) = \begin{cases} \bar{u}(t), & t \in [0, T], \\ -U^{-1}B^*e^{-A^*(t-T)}N^{-1}(1)\bar{x}(T), & t \in [T, T+1]. \end{cases}$$

It is clear that the control $\bar{\bar{u}}(t)$ transfers the point x_0 to the origin on the time interval $[0, T+1]$, and, because of (5.23) and (5.24), we have that

$$(5.26) \quad \int_0^{T+1} (U\bar{\bar{u}}(t), \bar{\bar{u}}(t)) dt < (Mx_0, x_0).$$

On the other hand, it is known that, for any control $\tilde{u}(t)$ transferring the point x_0 to the point 0 on the time interval $[0, T+1]$, the inequality

$$(5.27) \quad \int_0^{T+1} (U\tilde{u}(t), \tilde{u}(t)) dt \geq (N^{-1}(T+1)x_0, x_0) \geq (Mx_0, x_0)$$

is fulfilled.

The contradiction between (5.25) and (5.26) shows that $m(x_0) \geq (Mx_0, x_0)$. Considering (5.23), we infer that $m(x_0) = (Mx_0, x_0)$.

We now can prove assertion 1 of Theorem 5.3.

Let $x_0 \in L_-(\tilde{A})$ and let $x(t)$ be the solution of the equation with the control $u(x) = -U^{-1}B^*Mx$. Then $x(t) \rightarrow 0$ as $t \rightarrow \infty$. On the other hand, because of (5.13), we have that

$$(5.28) \quad \begin{aligned} \frac{d}{dt} (Mx(t), x(t)) &= (A^*Mx(t), x(t)) - 2(MBU^{-1}B^*Mx(t), x(t)) + (MAx(t), x(t)) \\ &= -(MBU^{-1}B^*Mx(t), x(t)) \\ &= -(Uu(x(t)), u(x(t))). \end{aligned}$$

Integrating (5.27) from 0 to T , we obtain that

$$\begin{aligned} \int_0^T (Uu(x(t)), u(x(t))) dt &= - \int_0^T \frac{d}{dt} (Mx(t), x(t)) dt \\ &= (Mx_0, x_0) - (Mx(T), x(T)) \rightarrow (Mx_0, x_0) \\ &= m(x_0) \quad \text{as } T \rightarrow \infty, \end{aligned}$$

which proves assertion 1 of Theorem 5.3.

Next, we turn to the proof of assertion 2 of the theorem. Suppose first that $x_0 \in L_0(\tilde{A}) = L_0(A)$ (recall Theorem 5.2). Because of Lemma 5.6, $L_0(A) \subset \text{Ker } M$, so that $m(x_0) = 0$. Since the control $u \equiv 0$ does not transfer the point x_0 to the point 0, assertion 2 is proved in this case.

Next, suppose that $x_0 \notin L_-(\tilde{A})$. Then, from Theorem 5.2, it follows that $x_0 = y_0 + z_0$, where $y_0 \in L_-(\tilde{A})$ and $z_0 \in L_0(\tilde{A})$. Let $v(t)$ be any control transferring the point x_0 to the point 0 asymptotically as $t \rightarrow \infty$ and let $x(t)$ be the solution of the equation corresponding to this control such that

$$J(v) = \int_0^\infty (Uv(t), v(t)) dt \geq m(x_0).$$

Denote by $y(t)$ the solution of (0.1) with initial condition y_0 and control $u(y) = -U^{-1}B^*My$. Because of the linearity of (0.1), it is clear that the control $w(t) = v(t) - u(y(t))$ transfers the point z_0 to the origin of coordinates asymptotically. A direct computation yields that

$$\begin{aligned} \frac{d}{dt} (My(t), x(t)) &= (M\tilde{A}y(t), x(t)) + (My(t), Ax(t) + Bv(t)) \\ &= ([MA - MBU^{-1}B^*M + A^*M]y(t), y(t)) + (My(t), Bv(t)). \end{aligned}$$

Because of (5.2), we obtain that

$$(5.29) \quad \begin{aligned} \frac{d}{dt} (My(t), x(t)) &= (My(t), Bv(t)) = (UU^{-1}B^*My(t), v(t)) \\ &= (Uu(y(t)), v(t)). \end{aligned}$$

Hence, for every $T > 0$, because of (5.27), we have that

$$\begin{aligned}
 & \int_0^T (Uw(t), w(t)) dt \\
 &= \int_0^T (Uv(t), v(t)) dt - 2 \int_0^T (Uv(t), u(y(t))) dt \\
 (5.30) \quad &+ \int_0^T (Uu(y(t)), u(y(t))) dt \\
 &= \int_0^T (Uv(t), v(t)) dt + 2(My(T), y(T)) \\
 &\quad - 2(My_0, x_0) + (My_0, y_0) - (My(T), y(T)) \\
 &\rightarrow (Mx_0, x_0) - 2(My_0, x_0) + (My_0, y_0) \quad \text{as } T \rightarrow \infty
 \end{aligned}$$

because $x(T)$, $y(T)$ tend to zero as $T \rightarrow \infty$. Theorem 5.2 implies that $Mx_0 = My_0$, so, from (5.29), we obtain that

$$J(w) = (My_0, x_0) - 2(My_0, x_0) + (Mx_0, y_0) = -(My_0, x_0) + (Mx_0, y_0) = 0$$

because $(Mx_0, y_0) = (My_0, x_0)$ is a real number. It follows that $W(t)$ must be identically equal to zero. However, as we have seen before, the zero control does not transfer the point z_0 to the point 0, so this contradiction proves assertion 2.

COROLLARY 5.4. *If the spectrum of the matrix A is disjoint from the imaginary axis, then, for every initial state $x_0 \in \mathbb{R}^n$, the control $u(x) = -U^{-1}B^*Mx$ gives the solution of the optimal stabilizability problem.*

Example. Let the controlled process be described by the system

$$\dot{x}_1 = -11x_1 + 6x_2, \quad \dot{x}_2 = -20x_1 + 11x_2 + u,$$

$$J(u) = \int_0^\infty u^2 dt,$$

$$A = \begin{pmatrix} -11 & 6 \\ -20 & 11 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Straightforward computations yield

$$\begin{aligned}
 N(\theta) &= \begin{pmatrix} \frac{9}{2}e^{2\theta} - \frac{9}{2}e^{-2\theta} - 18\theta & \frac{15}{2}e^{2\theta} - 9e^{-2\theta} - 33\theta + \frac{3}{2} \\ \frac{15}{2}e^{2\theta} - 9e^{-2\theta} - 33\theta + \frac{3}{2} & \frac{25}{2}e^{2\theta} - 18e^{-2\theta} - 60\theta + \frac{11}{2} \end{pmatrix}; \\
 N^{-1}(\theta) &= \begin{pmatrix} \frac{25}{2}e^{2\theta} - 18e^{-2\theta} - 60\theta + \frac{11}{2} & -\frac{15}{2}e^{2\theta} + \frac{9}{2}e^{-2\theta} + 33\theta - \frac{3}{2} \\ -\frac{15}{2}e^{2\theta} + \frac{9}{2}e^{-2\theta} + 33\theta - \frac{3}{2} & \frac{9}{2}e^{2\theta} - \frac{9}{2}e^{-2\theta} - 18\theta \end{pmatrix} \frac{1}{\frac{9}{4}e^{2\theta} + \frac{9}{4}e^{-2\theta} - 9\theta^2 - \frac{18}{4}}; \\
 M &= \begin{pmatrix} \frac{50}{9} & -\frac{10}{3} \\ -\frac{10}{3} & 2 \end{pmatrix}; \\
 u(x) &= -B^*Mx = \frac{10}{3}x_1 - 2x_2.
 \end{aligned}$$

With the control $u(x)$, the system takes the following form:

$$\begin{aligned}
 \dot{x}_1 &= -11x_1 + 6x_2, \quad \dot{x}_2 = -\frac{50}{3}x_1 + 9x_2. \\
 (\lambda_1 = \lambda_2 = -1).
 \end{aligned}$$

The optimal value of the cost function for the initial state $x = (x_1, x_2)$ is given by $m(x) = \frac{50}{9}x_1^2 + 2x_2^2 - \frac{20}{3}x_1x_2$.

REFERENCES

- [1] N. N. KRASOVSKY, *Problems of stabilization of controlled motions*, in Theory of Motion Stabilization, 2nd ed., I. G. Malkin, ed., Nauka, Moscow, 1966. (In Russian.)
- [2] L. M. LETOV, *Dynamics of Flight and Control*, Nauka, Moscow, 1969. (In Russian.)
- [3] E. B. LEE AND L. MARCUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [4] V. I. KOROBOV, *A general approach to the solution of the bounded control synthesis problem in a controllability problem*, Mat. Sb., 109(151) (1979), pp. 582–606; Math. USSR-Sb., 37 (1980). (English translation.)
- [5] V. I. KOROBOV AND G. M. SKLYAR, *On a set of positionally bounded controls solving the problem of synthesis*, Dokl. Akad. Nauk SSSR, 312 (1990), pp. 1304–1308; Soviet Math. Dokl., 41 (1990). (English translation.)
- [6] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.
- [7] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.
- [8] F. FLANDOLI, *Invertibility of Riccati operators and controllability of related systems*, Systems Control Lett., 9 (1987), pp. 65–72.
- [9] V. I. KOROBOV AND G. M. SKLYAR, *Time optimality and the power moment problem*, Mat. Sb., 134 (176) (1987), pp. 186–206; Math. USSR-Sb., 62 (1989), pp. 185–206. (English translation.)
- [10] M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

OPTIMALITY CONDITIONS VIA NORM SCALARIZATION IN VECTOR OPTIMIZATION*

PHAN QUOC KHANH†

Dedicated to Professor Stefan Rolewicz on his 60th birthday.

Abstract. An orthogonality concept in normed spaces is defined and used to obtain sufficient and necessary conditions for efficiency, weak efficiency, and proper efficiency via norm scalarization for the case of general (nonpointed) cones. As an application, a control approximation problem with a nonpointed ordering cone is considered.

Key words. norm scalarization, efficient points, weakly efficient points, properly efficient points, approximation

AMS(MOS) subject classifications. 90C31, 90C30, 90B50

1. Introduction. Scalarization means the replacement of the problem of finding an efficient (or weakly efficient, properly efficient) point of a subset S of the objective space X by a scalar problem of minimizing and/or maximizing a suitable functional f on X . If this functional is a norm, we have a norm scalarization. Perhaps this type of scalarization together with linear scalarization are of the most importance, because norms and linear functionals are in a sense simplest and best to handle. Furthermore, norm scalarization gives close relations between vector optimization and approximation theory.

Scalarization by the Euclidean norm was considered in [15], [21] among others, and by Chebyshev norm in [3]. For a complete review of various aspects of scalarization, including norm scalarization, in \mathbb{R}^n , the reader is referred to [19]. In particular, the weighted l_p norm, the weighted Chebyshev norm, and a composite norm were investigated there. Moreover, the mentioned scalarizing functional f is usually considered depending also on a vector parameter α , sometimes called a controlling parameter, since the decision maker can specify the values of α . A scalarizing functional f should desirably have certain properties. For instance, roughly speaking, we say that f is complete if and only if $(\hat{x}, \hat{\alpha})$ being a minimum of f (for some $\hat{\alpha}$, and on a suitable set of (x, α) related to S) is a necessary and sufficient condition for \hat{x} to be an efficient point of S . The completeness and various aspects of constructiveness of scalarization such as local controllability, robust computability, simplicity, and so on are considered in [19] for finite dimensions. Some of them are examined for differential games in [20].

The present paper concerns only one aspect: sufficient conditions and necessary conditions for efficient, weakly efficient, and properly efficient points via norm scalarization without a controlling parameter.

In [17] Wierzbicki proved the following sufficient condition. Assume that the ordering cone C of a Hilbert space X is such that

$$(1) \quad C \subseteq C^* := \{x^* \in X^* / \langle x^*, x \rangle \geq 0, \forall x \in C\}.$$

If \bar{x} is a unique minimum of $\|x - \hat{x}\|$ on S (or a unique maximum of it on a certain part of S), for a suitable reference point \hat{x} , then \bar{x} is an efficient point of S .

* Received by the editors April 4, 1990; accepted for publication (in revised form) July 23, 1991.

† Centre for Optimization and Control, I.P.O. Box 43, Hanoi, Vietnam.

Rolewicz [14] extended this result to normed spaces under the condition

$$(2) \quad C \cap (x - C) \subseteq B(O, \|x\|) \cup \{x\} \quad \text{for all } x \in X,$$

where $B(O, r)$ stands for the open ball of radius r and centered at O , and proved that if X is a Hilbert space, (1) and (2) are equivalent. In [7]–[9], Jahn explained (2) as the condition that the norm in X is strongly increasing and generalized the result to the case of general strongly increasing functionals, which need not be the norm of X . He proved various necessary conditions and sufficient conditions for efficient points, weakly efficient points, and properly efficient points in terms of norms additionally defined on X . Moreover, these scalarization results proved to have interesting applications in approximation theory [10]. However, all the mentioned conditions are much more restrictive than pointedness. In fact, (1) means that the angle between any two vectors of C must not be greater than $\pi/2$. In practice, ordering cones not satisfying this condition or even nonpointed are often met. For example, the lexicographic order does not fulfill (1), nor does any ordering cone, which is strictly greater than \mathbb{R}_+^n in the Euclidean space. But this situation is quite frequently faced because the preference of the decision maker may be “looser” than that defined by \mathbb{R}_+^n . For an example with nonpointed cone, see § 6.

In this paper, we extend these results to normed spaces ordered by general ordering cones. We prefer to consider scalarization without additional norms because this says more about the geometric structure of X , in addition to possible applications in approximation theory.

In § 2 we recall the definitions of the needed optimality notions in vector optimization and introduce an orthogonality concept in normed spaces. Section 3 is devoted to sufficient conditions under assumptions weaker than (1) and (2). We also explain how to apply the results. In the next section, similar results are proved for weak and proper efficiency. Section 5 deals with necessary conditions for efficiency and weak efficiency via norm scalarization. Finally, an application to a control approximation problem with a nonpointed ordering cone is examined in § 6.

2. Definitions. First recall some notions. Let S be a subset of a normed space X ordered by a convex cone C . A point $\bar{x} \in S$ is said to be an efficient point of S if there are no points $x \in S$ such that $x \in (\bar{x} - C) \setminus (\bar{x} + C)$. If the relative interior riC is nonempty and there is no $x \in S$ such that $x \in \bar{x} - riC$, \bar{x} is called a weakly efficient point of S . There are various notions of proper efficiency given by Kuhn and Tucker [13], Hurwicz [6], Geoffrion [4], Borwein [2], Benson [1], and Henig [5]. In the simplest case, where $X = \mathbb{R}^n$, $C = \mathbb{R}_+^n$, and S is convex, all these notions coincide. For a consideration and comparison of these notions in \mathbb{R}^n , see [16], and in infinite-dimensional spaces, see [11].

In the present paper, we use the following definition of Borwein. A point $\bar{x} \in S$ is said to be a properly efficient point of S if O is an efficient point of the tangent cone $T(S + C, \bar{x})$ of $S + C$ at \bar{x} . In the following, we will use the notation $E(S, C)$, $WE(S, C)$, and $PE(S, C)$ for the sets of all efficient points, weakly efficient points, and properly efficient points, respectively, of S (with respect to the ordering cone C).

We define an orthogonality in a normed space X as follows. Assume that X is a direct sum $X = L \oplus L^\perp$ of two closed subspaces, with $\text{codim } L > 1$. Then we say that L^\perp is orthogonal to L if the following monotonicity of the canonical projection $p: X \rightarrow L^\perp$ holds: $d(x_1, L) < d(x_2, L)$ implies $\|p(x_1)\| < \|p(x_2)\|$. Observe that if L^\perp is orthogonal to L , then $d(x_1, L) = d(x_2, L)$ implies $\|p(x_1)\| = \|p(x_2)\|$. Indeed, $d(tx_1, L) < d(x_2, L)$ for $t \in (0, 1)$. Hence, $\|p(tx_1)\| < \|p(x_2)\|$. Taking the limit as $t \rightarrow 1$, we obtain $\|p(x_1)\| \leq \|p(x_2)\|$. Similarly, $\|p(x_2)\| \leq \|p(x_1)\|$.

Note that if $\text{codim } L = 1$, the above monotonicity always holds and then does not specify an orthogonality.

PROPOSITION 2.1. *Let X be a Hilbert space and L be a closed subspace. Then the orthogonal complement L^\perp (in the usual sense) is orthogonal to L . Conversely, if L^\perp is orthogonal to L and $L^+ \cap L^\perp = \{0\}$, then $L^+ = L^\perp$.*

Proof. L^\perp is clearly orthogonal to L . Now let L^+ be orthogonal to L and $L^+ \cap L^\perp \neq \{0\}$. Suppose $x_1 \in L^+ \setminus L^\perp$ exists. Then $d(x_1, L) < \|x_1\|$. We find $l \in L$ satisfying $d(x_1, L) < \|x_1 - l\| < \|x_1\|$. In $L^+ \cap L^\perp \neq \{0\}$, there exists x_2 with $\|x_2\| = \|x_1 - l\|$. Hence, $d(x_2, L) = \|x_2\| > d(x_1, L)$, which contradicts the fact that $\|p(x_2)\| = \|x_2\| < \|x_1\| = \|p(x_1)\|$. So, $L^+ \subseteq L^\perp$. If L^+ was a proper subset of L^\perp , there would exist $0 \neq x \in L^\perp$ such that $\langle x, l \rangle = 0$ for all $l \in L^+$. This and the fact that $\langle x, l \rangle = 0$ for all $l' \in L$ together imply that $\langle x, y \rangle = 0$ for all $y \in L \oplus L^+ = X$. This contradiction shows that $L^+ = L^\perp$. \square

Note that if X is not a Hilbert space, there may not exist L^+ orthogonal to a given closed subspace L . The following example shows that our orthogonality assumption imposed in the subsequent theorems is weaker than the usual orthogonality if X is a Hilbert space. It also ensures that the assumption $L^+ \cap L^\perp \neq \{0\}$ in Proposition 2.1 is essential.

Example 2.2 (Penot, private communication). Let X be the Euclidean space \mathbb{R}^4 , $L = \{(0, x_2, x_3, 0) \in X / x_2, x_3 \in \mathbb{R}\}$. Then, both $L^\perp = \{(x_1, 0, 0, x_4) \in X / x_1, x_4 \in \mathbb{R}\}$ and $L^+ = \{x \in X / x_1 = x_2, x_3 = x_4\}$ are orthogonal to L . Note that $L^+ \cap L^\perp = \{0\}$.

In what follows, for a set $A \subset X$, by A^+ we denote the projection $p(A)$ on L^+ .

3. Sufficient conditions for efficiency. Throughout this paper, let X be a normed space ordered by a convex cone C with $\text{cl } C \neq X$ and let S be a nonempty subset of X .

THEOREM 3.1. *Assume that $X = L \oplus L^+$, where L^+ is orthogonal to L if $\text{codim } L > 1$. Assume further that, for all $x_1, x_2 \in C^+$,*

$$(3) \quad \|\alpha x_1 + \beta x_2\| \leq \|x_1 + x_2\|$$

whenever $\alpha, \beta \in [0, 1]$. Then the following assertions hold:

(a) *If $S \subseteq \hat{x} + C$ for some $\hat{x} \in X$, then any point $\bar{x} \in X$ satisfying*

$$(4) \quad d(\bar{x}, \hat{x} + L) < d(x, \hat{x} + L) \quad \text{for all } x \in S \setminus \{\bar{x}\}$$

is an efficient point of S ;

(b) *Every point $\bar{x} \in (\tilde{x} - C) \cap S$, for some $\tilde{x} \in S$, satisfying*

$$(5) \quad d(\bar{x}, \tilde{x} + L) > d(x, \tilde{x} + L) \quad \text{for all } x \in ((\tilde{x} - C) \cap S) \setminus \{\bar{x}\}$$

is an efficient point of S .

Proof. (a) We first show that $p(\bar{x}) \in E(S^+, C^+)$. Since $S \subseteq \hat{x} + C$, $S^+ \subseteq p(\hat{x}) + C^+$. Then $(p(\bar{x}) - C^+) \cap S^+ = (p(\bar{x}) - C^+) \cap S^+ \cap (p(\hat{x}) + C^+)$. Let z be any point in the latter set. Then $z - p(\hat{x}) \in C^+$ and $p(\bar{x}) - p(\hat{x}) - (z - p(\hat{x})) \in C^+$. Therefore by (3),

$$\|z - p(\hat{x})\| \leq \|p(\bar{x}) - p(\hat{x})\|.$$

Since L^+ is orthogonal to L , (4) implies that

$$\|p(\bar{x}) - p(\hat{x})\| < \|z - p(\hat{x})\| \quad \text{for all } z \in S^+ \setminus \{p(\bar{x})\}.$$

Thus $(p(\bar{x}) - C^+) \cap S^+ = \{p(\bar{x})\}$, i.e., $p(\bar{x}) \in E(S^+, C^+)$.

Now suppose $\bar{x} \notin E(S, C)$, i.e., there exists $x \in S$ such that $x \in (\bar{x} - C) \setminus (\bar{x} + C)$. Then, $p(x) \in p(\bar{x}) - C^+$. Assertion (a) will be proved if we show that $p(x) \notin p(\bar{x}) + C^+$, because this contradicts the efficiency of $p(\bar{x})$. Note that (3) implies that C^+ is pointed, for if there exists $x_1 \in C^+ \setminus \{0\}$ such that $x_2 = -x_1 \in C^+$, then, taking $\alpha = 1, \beta = 0$, we arrive at a contradiction $\|x_1\| \leq 0$. So if $p(x) \in p(\bar{x}) + C^+$, we have $p(x) = p(\bar{x})$. Then by the orthogonality, $d(x, \hat{x} + L) = d(\bar{x}, \hat{x} + L)$, contradicting (4).

(b) Since \bar{x} is the center of the straight interval $(2\bar{x} - \tilde{x}, \tilde{x})$, by (5) and by the symmetry of the balls in X we have, for all $x \in ((\tilde{x} - C) \cap S) \setminus \{\bar{x}\}$,

$$d(\bar{x}, 2\bar{x} - \tilde{x} + L) < d(x, 2\bar{x} - \tilde{x} + L).$$

Observing that $\bar{x} \in (\tilde{x} - C) \cap S \cap (2\bar{x} - \tilde{x} + C) \subseteq 2\bar{x} - \tilde{x} + C$, we can apply assertion (a) to see that $\bar{x} \in E((\tilde{x} - C) \cap S \cap (2\bar{x} - \tilde{x} + C), C)$. We claim that $\bar{x} \in E((\tilde{x} - C) \cap S, C)$. Indeed, suppose there exists $x' \in (\tilde{x} - C) \cap S \cap ((\bar{x} - C) \setminus (\bar{x} + C))$. Then $p(\bar{x}) - p(x') \in C^+$. Since $p(\tilde{x}) - p(\bar{x}) \in C^+$, by (3),

$$\|p(\tilde{x}) - p(\bar{x})\| \leq \|p(\tilde{x}) - p(\bar{x}) + p(\bar{x}) - p(x')\| = \|p(\tilde{x}) - p(x')\|.$$

This and (5) together show that $x' = \bar{x}$, which is a contradiction. Thus $\bar{x} \in E((\tilde{x} - C) \cap S, C)$, i.e., $(\bar{x} - C) \cap (\tilde{x} - C) \cap S \subseteq \bar{x} + C$.

Finally, as $\bar{x} - C \subseteq \tilde{x} - C$, the last inclusion implies

$$(\bar{x} - C) \cap S = (\bar{x} - C) \cap (\tilde{x} - C) \cap S \subseteq \bar{x} + C,$$

i.e., $\bar{x} \in E(S, C)$. \square

Note that it is rather restrictive to assume that $S \subseteq \hat{x} + C$. Therefore, assertion (b), being true without this assumption, is important. Moreover we can ask the natural question of whether the theorem remains true if we strengthen (3) to the strict inequality and at the same time weaken (4) and (5), allowing nonstrict inequalities to hold. This is indeed the case, but under an additional assumption as follows.

THEOREM 3.1'. *Under the additional assumption that $L = \text{cl } C \cap (-\text{cl } C)$, Theorem 3.1 is still true if (3) is replaced by*

$$(3') \quad \|\alpha x_1 + \beta x_2\| < \|x_1 + x_2\|$$

for all $x_1, x_2 \in C^+$ such that $x_1 + x_2 \neq 0$ and all $\alpha, \beta \in [0, 1]$ with $\alpha + \beta < 2$, and if (4) and (5) are replaced by the nonstrict inequalities.

Proof. (a) Similarly as above, we have $p(\bar{x}) \in E(S^+, C^+)$. Suppose $\bar{x} \notin E(S, C)$, i.e., there exists $x \in S$ such that $x \in (\bar{x} - C) \setminus (\bar{x} + C)$. As before, if $p(x) \in (p(\bar{x}) - C^+) \cap (p(\bar{x}) + C^+)$, then $p(x) = p(\bar{x})$. Since $L = \text{cl } C \cap (-\text{cl } C)$, $x \in (\bar{x} - \text{cl } C) \cap (\bar{x} + \text{cl } C)$. This contradicts the fact that $x \notin \bar{x} + C$.

(b) This assertion is proved by applying (a) as for Theorem 3.1. \square

The following examples confirm that the assumptions of the theorems are essential.

Example 3.2 (the orthogonality is essential). Let X be the Euclidean space \mathbb{R}^3 . Assume, for Fig. 1, that $|\overline{ob}| = |\overline{oc}|$ and $|\overline{ab}| = |\overline{ac}|$ (by $|\cdot|$ we denote the length of a straight interval), and that $\angle oab = \angle oac = (\pi/2)$. Let the cone C be the part of \mathbb{R}^3 , defined by two planes oab and oac , corresponding to the angle bac . Let $\overline{de} \parallel \overline{ac}$. Let d be so near to b that $|\overline{ad}| > |\overline{ae}|$ and that $\angle aed > (\pi/2)$ (since $\angle bac > \pi/2$, see Fig. 1(b)). Let S be the interval \overline{de} . Let L be the plane oxy , which is not orthogonal to $L = C \cap (-C)$. Then all other assumptions of Theorems 3.1(a) and 3.1'(a) are satisfied with $\bar{x} = e$, since

$$d(e, L) < d(u, L) \quad \text{for all } u \in \overline{de}, u \neq e,$$

but $e \notin E(S, C)$ (d is the unique efficient point of S).

Similarly, we can show that the orthogonality is essential for assertion (b) (for instance, we can easily construct a similar counterexample in \mathbb{R}^3 with the ordering cone being $-C$, where C is defined in Example 3.2).

Example 3.3 (assumption that $L = \text{cl } C \cap (-\text{cl } C)$ is essential). Let X be the Euclidean space \mathbb{R}^3 ordered by \mathbb{R}_+^3 . Let $S = \{(0, 0, z) \mid z \in (0, 1]\}$. Then $E(S, C) = \emptyset$. Let L be the straight line \overline{oz} and L^+ be the plane oxy . Then, except for the assumption that $L = \text{cl } C \cap (-\text{cl } C)$, all assumptions of Theorem 3.1' are satisfied for every point of S , although none are efficient.

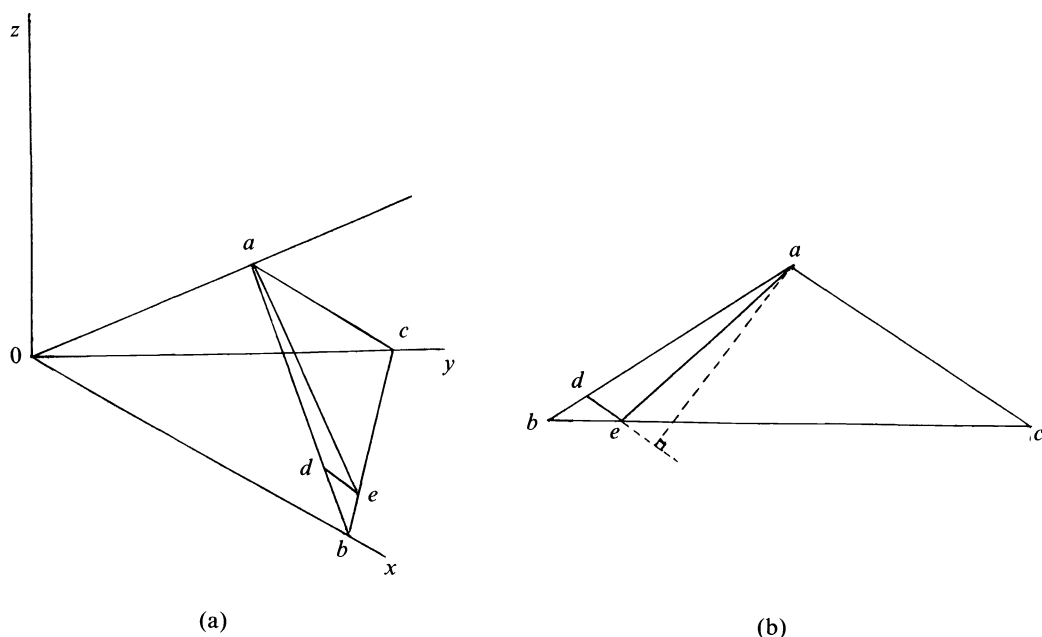


FIG. 1

The following lemma assures us that Theorems 3.1 and 3.1' are extensions of the corresponding results of Wierzbicki, Rolewicz, and Jahn.

LEMMA 3.4. *The following conditions are equivalent in pairs, (i)–(i') and (ii)–(ii'):*

- (i) Condition (3) holds;
- (ii) Condition (3') holds;
- (i') $C^+ \cap (x - C^+) \subseteq L^+ \cap \text{cl } B(0, \|x\|)$ for all $x \in L^+$;
- (ii') $C^+ \cap (x - C^+) \subseteq L^+ \cap B(0, \|x\|) \cup \{x\}$ for all $x \in L^+$.

Proof. We prove only the equivalence of (ii) and (ii'). The equivalence of (i) and (i') is proved similarly. Suppose there exist $x_1, x_2 \in C^+$ and $\alpha, \beta \in [0, 1]$ with $\alpha + \beta < 2$ such that $\|\alpha x_1 + \beta x_2\| \geq \|x_1 + x_2\|$. Put $x = x_1 + x_2$ and $z = \alpha x_1 + \beta x_2$. Then $z \in C^+ \cap (x - C^+)$, but $z \notin L^+ \cap B(0, \|x\|) \cup \{x\}$.

Conversely, if $z \in C^+ \cap (x - C^+)$, then $x - z \in C^+$. By (ii), if $x - z \neq 0$, we have

$$\|z\| < \|z + x - z\| = \|x\|.$$

Thus, (ii') holds. \square

Of course, we can restate (i) and (ii) in terms of increasing norms following Jahn to see clearer the relations between the above theorems and the corresponding results of Jahn [7], [9]. Note also that sometimes a statement of the form (i') and (ii') is more convenient than (i) and (ii), as in the case of Theorem 4.1 below. Furthermore, it may be interesting that in the case of Hilbert spaces, (i) and (ii) are equivalent although in general (ii) is stronger than (i). To prove this, we can show that (i') implies that $C^+ \subseteq (C^+)^*$, where the conjugation is considered on L^+ , in the same way as in Theorem 3 of [14].

We should point out the following important fact concerning possibilities of applying Theorems 3.1 and 3.1'. Both conditions (i) and (ii) are satisfied when L^+ is one-dimensional. So assertions (a) and (b) of these theorems always hold if L is any closed hyperplane of the normed space X if C is contained in a closed half-space defined by P . Moreover, since $\text{cl } C \neq X$ and C is a convex cone, such a P always

exists. In general, there may be many subspaces L satisfying the assumptions of these theorems. However, since $d(\bar{x}, \hat{x} + L) = \inf_{l \in L} \|\bar{x} - \hat{x} + l\|$, the bigger is L , the more difficult is the problem of finding \bar{x} . The following propositions are helpful in looking for a space L as small as possible.

PROPOSITION 3.5. *Let X be a Hilbert space. Let an orthogonal sum $L = L_1 \oplus L_2$ be a subspace of X . Let p_1 be the orthogonal projection of X onto L_1 . Assume that L satisfies (3) and that $\langle x, x' \rangle \geq 0$ for all $x, x' \in p_1(C)$. Then L_1 should satisfy (3) as well.*

Proof. Let p and \tilde{p} be the orthogonal projections of X onto L^\perp and L_2^\perp , respectively. Then for arbitrary c, c' in C we have

$$\begin{aligned} \langle \tilde{p}(c), \tilde{p}(c') \rangle &= \langle p(c) + p_1(c), p(c') + p_1(c') \rangle \\ &= \langle p(c), p(c') \rangle + \langle p_1(c), p_1(c') \rangle. \end{aligned}$$

Since L satisfies (3), $p(C) \subseteq (p(C))^*$, where the conjugation is considered in L^\perp . Hence $\langle p(c), p(c') \rangle \geq 0$. Thus $\langle \tilde{p}(c), \tilde{p}(c') \rangle \geq 0$, i.e., L_2 fulfils (3). \square

Note further that each subspace L satisfying (3) should contain $L_0 = \text{cl } C \cap (-\text{cl } C)$. Therefore a question may arise: can we extend L_0 to a subspace L satisfying (3)? The following simple result gives an answer.

PROPOSITION 3.6. *Let X be a Hilbert space. Let $L_0 = \text{cl } C \cap (-\text{cl } C)$. Let H be a set of all possible h in L_0^\perp meeting the following conditions:*

- (i) *All points of H are orthogonal in pairs;*
 - (ii) *For each $h \in H$, there exist $c, c' \in C$ such that $h = p(c)$ and $\langle p(c), p(c') \rangle < 0$, where p is the orthogonal projection of X onto $(L_0 \oplus (H \setminus \{h\}))^\perp$.*
- Then, the subspace $L = L_0 \oplus H$ satisfies (3).*

Proof. Suppose, to the contrary, that there exist $c_1, c_2 \in C$ such that $\langle \bar{p}(c_1), \bar{p}(c_2) \rangle < 0$, where \bar{p} stands for the orthogonal projections of X onto $L_0 \oplus H$. Then $H_1 = H \cup \{\bar{p}(c_1)\}$ also satisfies (i) and (ii). This contradicts the definition of H . \square

In practice, especially when X is a Euclidean space, we can expect that by applying Proposition 3.6 a possibly small subspace L meeting condition (3) may be constructed, starting with L_0 as follows. Let p be the orthogonal projection onto L_0^\perp . If $\langle p(c), p(c') \rangle \geq 0$ for all $c, c' \in C$, L_0 clearly satisfies (3). If there are $c_1, c_2 \in C$ with $\langle p(c_1), p(c_2) \rangle < 0$, we put $h_1 = p(c_1)$. It is added here that, since $L_0 = \text{cl } C \cap (-\text{cl } C)$, $p(\text{cl } C) = L_0^\perp \cap \text{cl } C$. Hence, checking the negativity of the above inner products is rather simple. Next, denoting by p_1 the orthogonal projection of X onto $(L_0 + \{\gamma h_1 : \gamma \in \mathbb{R}\})^\perp$, we check whether there are $c'_1, c'_2 \in C$ such that $\langle p_1(c'_1), p_1(c'_2) \rangle < 0$. If there are no such c'_1 and c'_2 , $L_0 + \{\gamma h_1 : \gamma \in \mathbb{R}\}$ is a desired subspace L . If otherwise, we continue to do similarly.

After this thorough discussion of the significance of Theorems 3.1 and 3.1' for efficiency, we are convinced that similar results for weak and proper efficiency are important. Such results are the content of the next section.

4. Sufficient conditions for weak and proper efficiency.

THEOREM 4.1. *Assume that $X = L \oplus L^+$, where L^+ is orthogonal to L if $\text{codim } L > 1$, and that $\text{ri } C \neq \emptyset$. Assume further that*

$$(6) \quad C^+ \cap (u - \text{ri } C^+) \subseteq L^+ \cap B(O, \|u\|) \quad \text{for all } u \in L^+.$$

Then the following assertions hold:

- (a) *If $S \subseteq \hat{x} + C$ for some $\hat{x} \in X$, then any point $\bar{x} \in S$ satisfying*

$$(7) \quad d(\bar{x}, \hat{x} + L) \leq d(x, \hat{x} + L) \quad \text{for all } x \in S$$

is a weakly efficient point of S ;

(b) Every point $\bar{x} \in (\tilde{x} - C) \cap S$, for some $\tilde{x} \in S$, satisfying

$$(8) \quad d(\bar{x}, \tilde{x} + L) \geq d(x, \tilde{x} + L) \quad \text{for all } x \in (\tilde{x} - C) \cap S$$

is a weakly efficient point of S .

Proof. (a) It follows from (6) that

$$(p(\hat{x}) + C^+) \cap (p(\bar{x}) - \text{ri } C^+) \subseteq L^+ \cap B(p(\hat{x}), \|p(\bar{x}) - p(\hat{x})\|).$$

On the other hand, the orthogonality and (7) together imply that

$$\|p(\bar{x}) - p(\hat{x})\| \leq \|p(x) - p(\bar{x})\| \quad \text{for all } x \in S.$$

Consequently, since $S^+ \subset p(\hat{x}) + C^+$, $S^+ \cap (p(\bar{x}) - \text{ri } C^+)$ is empty, i.e., $p(\bar{x}) \in WE(S^+, C^+)$.

To show that $\bar{x} \in WE(S, C)$, suppose there exists $x \in S \cap (\bar{x} - \text{ri } C)$. Then, $p(x) \in (S \cap (\bar{x} - \text{ri } C))^+ \subset S^+ \cap (p(\bar{x}) - \text{ri } C^+)$. This contraposition completes the proof of (a).

(b) By the same argument as in the proof of Theorem 3.1, we obtain that $\bar{x} \in WE((\tilde{x} - C) \cap S \cap (2\bar{x} - \tilde{x} + C), C)$. Furthermore, $\bar{x} \in WE((\tilde{x} - C) \cap S, C)$. For, if there exists $x' \in (\tilde{x} - C) \cap S \cap (\bar{x} - \text{ri } C)$, then $p(x') \in ((\tilde{x} - C) \cap S)^+ \cap (p(\bar{x}) - \text{ri } C^+)$. By (6) and by the symmetry of the balls in L^+ , $\|p(x') - p(\hat{x})\| > \|p(\bar{x}) - p(\tilde{x})\|$. This together with the orthogonality imply that $d(x', \tilde{x} + L) > d(\bar{x}, \tilde{x} + L)$. So $x' \notin (\tilde{x} - C) \cap S$ in view of (8). This contradiction confirms that $\bar{x} \in WE((\tilde{x} - C) \cap S, C)$. Now, observing that

$$(\bar{x} - \text{ri } C) \subseteq \bar{x} - C \subseteq \tilde{x} - C,$$

we obtain

$$(\bar{x} - \text{ri } C) \cap S = (\bar{x} - \text{ri } C) \cap S \cap (\tilde{x} - C) = \emptyset,$$

which allows us to conclude that $\bar{x} \in WE(S, C)$ \square

Remark 4.2. If $\text{ri } C^+ \neq \emptyset$, it is not hard to see the following connections between condition (6) and conditions (3), (3'): (3') \Rightarrow (6) \Rightarrow (3). If X is a Hilbert space, all these conditions are equivalent. But in the case of normed spaces, we can easily find examples showing that the converse implications are not true. Theorem 4.1(a) is an extension of the corresponding result of Jahn [7].

In the following, by $\text{reco } C$ ($\text{cor } C$) we denote the relative algebraic interior of C (the algebraic interior of C , respectively).

THEOREM 4.3. Let $X = L \oplus L^+$, where L^+ is orthogonal to L if $\text{codim } L > 1$. Assume that

$$(9) \quad C^+ \cap (x - C^+) \subseteq L^+ \cap B(O, \|x\|) \cup \{x\} \quad \text{for all } x \in L^+.$$

Then the following assertions hold:

(a) If $\text{reco } C \neq \emptyset$ and if $S \subseteq \hat{x} + C$ for some $\hat{x} \in X$, then any point $\bar{x} \in \hat{x} + \text{reco } C$ such that

$$(10) \quad d(\bar{x}, \hat{x} + L) \leq d(x, \hat{x} + L) \quad \text{for all } x \in S$$

is a properly efficient point of S ;

(b) If $\text{int } C \neq \emptyset$, then every point $\bar{x} \in (\tilde{x} - \text{int } C) \cap S$ for some $\tilde{x} \in S$ such that

$$(11) \quad d(\bar{x}, \tilde{x} + L) \geq d(x, \tilde{x} + L) \quad \text{for all } x \in (\tilde{x} - \text{int } C) \cap S$$

is a properly efficient point of S .

Proof. (a) First, we claim that

$$(12) \quad d(\bar{x}, \hat{x} + L) \leq d(x, \hat{x} + L) \quad \text{for all } x \in S + C.$$

Indeed, suppose there exist $x_1 \in S$ and $c_1 \in C$ such that

$$d(x_1 + c_1, \hat{x} + L) < d(\bar{x}, \hat{x} + L).$$

Then

$$\|p(x_1) + p(c_1) - p(\hat{x})\| < \|p(\bar{x}) - p(\hat{x})\|.$$

Since $p(c_1) \in C^+$ and $p(x_1) - p(\hat{x}) \in C^+$, we have

$$p(x_1) - p(\hat{x}) \in C^+ \cap (p(x_1) - p(\hat{x}) + p(c_1) - C^+).$$

Hence by (9),

$$\|p(x_1) - p(\hat{x})\| \leq \|p(x_1) + p(c_1) - p(\hat{x})\| < \|p(\bar{x}) - p(\hat{x})\|,$$

which contradicts (10). Thus (12) holds.

It is easy to verify that $d(\cdot, \hat{x} + L)$ is convex and continuous. Therefore (12) implies (see, e.g., [12, p. 156])

$$(13) \quad d(\bar{x}, \hat{x} + L) \leq d(\bar{x} + h, \hat{x} + L) \quad \text{for all } h \in T(S + C, \bar{x}).$$

Clearly, (13) also holds for all $h \in (\hat{x} - \bar{x} + C) \cap T(S + C, \bar{x})$. As $h \in \hat{x} - \bar{x} + C$, i.e., $\bar{x} + h \in \hat{x} + C$, we can apply Theorem 3.1' to obtain that

$$(14) \quad O \in E((\hat{x} - \bar{x} + C) \cap T(S + C, \bar{x}), C).$$

Suppose now that $\bar{x} \notin PE(S, C)$, i.e., $O \notin E(T(S + C, \bar{x}), C)$. Then there exists a nonzero vector $x \in ((-C) \setminus (C)) \cap T(S + C, \bar{x})$. Since $\bar{x} \in \hat{x} + \text{recor } C$, for sufficiently small $\gamma > 0$ we have $\bar{x} + \gamma x \in x + C$. Therefore

$$\gamma x \in (\hat{x} - \bar{x} + C) \cap ((-C) \setminus (C)) \cap T(S + C, \bar{x}).$$

This contradiction to (14) leads to assertion (a).

(b) The symmetry and (11) together imply that

$$d(\bar{x}, 2\bar{x} - \tilde{x} + L) \leq d(x, 2\bar{x} - \tilde{x} + L)$$

for all $x \in (\tilde{x} - \text{int } C) \cap S \cap (2\bar{x} - \tilde{x} + C) \subseteq 2\bar{x} - \tilde{x} + C$. Since $\bar{x} \in \tilde{x} - \text{int } C$, $\bar{x} \in 2\bar{x} - \tilde{x} + \text{int } C$ by the symmetry. According to assertion (a), $\bar{x} \in PE((\tilde{x} - \text{int } C) \cap S \cap (2\bar{x} - \tilde{x} + C), C)$, or, what is the same, $O \in E(T((\tilde{x} - \text{int } C) \cap S \cap (2\bar{x} - \tilde{x} + C) + C, \bar{x}), C)$. Now assertion (b) will follow immediately if we show that

$$(15) \quad T((\tilde{x} - \text{int } C) \cap S \cap (2\bar{x} - \tilde{x} + C) + C, \bar{x}) = T(S + C, \bar{x}).$$

Noting that $\bar{x} \in \text{int}((\tilde{x} - \text{int } C) \cap (2\bar{x} - \tilde{x} + C))$ we find a neighbourhood $N(\bar{x}) \subseteq (\tilde{x} - \text{int } C) \cap (2\bar{x} - \tilde{x} + C)$. Then

$$N(\bar{x}) \cap ((\tilde{x} - \text{int } C) \cap S \cap (2\bar{x} - \tilde{x} + C) + C) = N(\bar{x}) \cap (S + C).$$

Observing that the tangent cone of a set at a point depends only on the points of the set in a neighborhood of the given point, we arrive at (15). \square

Remark 4.4. Assertion (a) is an extension of Theorem 2.2 in [8], which corresponds to the case $L^+ = X$, $L = \{0\}$, $\text{cor } C \neq \emptyset$ and $S \subseteq \hat{x} + \text{cor } C$. In that theorem it is assumed that some conditions, which are more restrictive than (9) and (10), are satisfied for an additional norm in X . This norm is assumed to be not stronger than the norm of X . The concluded properness of \bar{x} is stated with respect to the original norm. However, we observe that the tangent cone corresponding to the additional norm $T^{\text{ad}}(S + C, \bar{x})$ contains $T(S + C, \bar{x})$. Hence $O \in E(T^{\text{ad}}(S + C, \bar{x}), C)$ implies $O \in E(T(S + C, \bar{x}), C)$. Thus using an additional norm does not make the theorem stronger.

5. Necessary conditions for efficiency and weak efficiency. In general, the above-formulated sufficient conditions are not necessary conditions, as we can easily find counterexamples. However, we have the following related necessary conditions.

THEOREM 5.1. *Let X be a normed space ordered by a closed convex cone C . Assume that $X = L \oplus L^+$, where $L = C \cap (-C)$ and L^+ is orthogonal to L if $\text{codim } L > 1$. Assume further that $\text{cor } C^+$ is nonempty. Then the following assertions hold.*

(a) *Each efficient point \bar{x} of an arbitrary set S satisfies the following condition: For a given \hat{x} such that $p(\hat{x}) \in p(\bar{x}) - \text{cor } C^+$,*

$$d(\bar{x}, \hat{x} + L) \leq d(x, \hat{x} + L)$$

and

$$\|p(\bar{x}) - p(\hat{x})\| < \|p(x) - p(\hat{x})\|$$

whenever $p(x) \neq p(\bar{x})$, for all $x \in S$ if and only if the norm (of X) considered on L^+ is the Minkowski functional corresponding to the order interval $[p(\hat{x}) - p(\bar{x}), p(\bar{x}) - p(\hat{x})]$, i.e.,

$$(16) \quad L^+ \cap \text{cl } B(O, \|p(\bar{x}) - p(\hat{x})\|) = (p(\hat{x}) - p(\bar{x}) + C^+) \cap (p(\bar{x}) - p(\hat{x}) - C^+).$$

(b) *Each efficient point \bar{x} of an arbitrary set S satisfies the condition: For a given \tilde{x} such that $p(\tilde{x}) \in p(\bar{x}) + \text{cor } C^+$,*

$$(17) \quad d(\bar{x}, \tilde{x} + L) \geq d(x, \tilde{x} + L)$$

and

$$(18) \quad \|p(\bar{x}) - p(\tilde{x})\| > \|p(x) - p(\tilde{x})\|$$

whenever $p(x) \neq p(\tilde{x})$, for all $x \in (\tilde{x} - C) \cap S$, if and only if the norm considered on L^+ is the Minkowski functional corresponding to the order interval $[p(\bar{x}) - p(\tilde{x}), p(\tilde{x}) - p(\bar{x})]$.

Proof. (a) Assume that (16) holds. Arguing by contradiction, suppose there exists $x' \in S$ with $\|p(x') - p(\hat{x})\| \leq \|p(\bar{x}) - p(\hat{x})\|$. Then $p(x') \in [2p(\hat{x}) - p(\bar{x}), p(\bar{x})]$. Hence $p(x') \in p(\bar{x}) - C^+$. If $p(x') \in p(\bar{x}) + C^+$, we have, by the pointedness of C^+ , $p(\bar{x}) = p(x')$. If $p(x') \notin p(\bar{x}) + C^+$, i.e.,

$$(19) \quad p(x') \in (p(\bar{x}) - C^+) \setminus (p(\bar{x}) + C^+),$$

then $x' \notin \bar{x} + C$. Moreover, $x' \in x - C$. Indeed, suppose $x' \in \bar{x} - C$. Then, since $L = C \cap (-C)$, $(x' + L) \cap (\bar{x} - C) = \emptyset$, for if $x' + l \in \bar{x} - C$ for some $l \in L$ we would have $x' = x' + l - l \in \bar{x} - C - l \subset \bar{x} - C$. Therefore, observing that $p(C) = C \cap L^+$, $p(x' + L) = p(x')$, we derive that $\{p(x')\} \cap (p(\bar{x}) - C^+) = \emptyset$. This contradiction to (19) asserts that $x' \in \bar{x} - C$. This in turn is impossible due to the efficiency of \bar{x} .

Now suppose there exists $x' \in S$ with $d(x', \hat{x} + L) < d(\bar{x}, \hat{x} + L)$. Then $\|p(x') - p(\hat{x})\| < \|p(\bar{x}) - p(\hat{x})\|$. As in the preceding part of the proof, this leads to a contradiction.

Conversely suppose (16) does not hold. Then at least one of the following two possibilities must occur.

(i) There exists $u \in L^+$ satisfying

$$\|u - p(\hat{x})\| < \|p(\bar{x}) - p(\hat{x})\|, \quad u \notin p(\bar{x}) - C^+.$$

Then the set S consisting of \bar{x} and any point x of $p^{-1}(u)$ has \bar{x} as an efficient point, but $d(x, \hat{x} + L) < d(\bar{x}, \hat{x} + L)$.

(ii) There exists $u \in L^+$ satisfying

$$\|p(\bar{x}) - p(\hat{x})\| < \|u - p(\hat{x})\|, \quad u \in (p(\bar{x}) - C^+) \setminus (p(\bar{x}) + C^+).$$

Then for the set S consisting of \bar{x} and any point x of $p^{-1}(u)$ we have $x \in E(S, C)$, but $d(\bar{x}, \hat{x} + L) < d(x, \hat{x} + L)$.

(b) We observe that $p(2\bar{x} - \tilde{x}) \in p(\bar{x}) - \text{cor } C^+$ and that

$$[p(\bar{x}) - p(\tilde{x}), p(\tilde{x}) - p(\bar{x})] = [p(2\bar{x} - \tilde{x}) - p(\bar{x}), p(\bar{x}) - p(2\bar{x} - \tilde{x})].$$

Furthermore, $\bar{x} \in E((\tilde{x} - C) \cap S, C)$ whenever $\bar{x} \in E(S, C)$.

These facts allow us to apply assertion (a) to obtain that

$$d(\bar{x}, 2\bar{x} - \tilde{x} + L) \leq d(x, 2\bar{x} - \tilde{x} + L)$$

and

$$\|p(\bar{x}) - p(2\bar{x} - \tilde{x})\| < \|p(x) - p(2\bar{x} - \tilde{x})\|$$

whenever $p(x) \neq p(\bar{x})$, for all $x \in (\tilde{x} - C) \cap S$. This, by virtue of the symmetry of the balls, implies that

$$d(\bar{x}, \tilde{x} + L) \geq d(x, \tilde{x} + L), \quad \|p(\bar{x}) - p(\tilde{x})\| > \|p(x) - p(\tilde{x})\|$$

whenever $p(x) \neq p(\bar{x})$, for all $x \in (\tilde{x} + C) \cap S$.

Conversely, if (17) and (18) hold, then by the symmetry of the balls we have

$$d(\bar{x}, 2\bar{x} - \tilde{x} + L) \leq d(x, 2\bar{x} - \tilde{x} + L)$$

and

$$\|p(\bar{x}) - p(2\bar{x} - \tilde{x})\| < \|p(x) - p(2\bar{x} - \tilde{x})\|$$

whenever $p(x) \neq p(\bar{x})$, for all $x \in (\tilde{x} - C) \cap S$. So, we can apply assertion (a) to see that the norm on L^+ is the Minkowski functional corresponding to the order interval

$$[p(2\bar{x} - \tilde{x}) - p(\bar{x}), p(\bar{x}) - p(2\bar{x} - \tilde{x})] = [p(\bar{x}) - p(\tilde{x}), p(\tilde{x}) - p(\bar{x})].$$

Theorem 5.1(a) extends Theorem 3.6 of [7]. The following example explains why the assumption that $L = C \cap (-C)$ is essential.

Example 5.2. Let $X = \mathbb{R}^3$ with the norm defined as follows

$$\|(x, y, z)\| = ((\max\{|x|, |y|\})^2 + |z|^2)^{1/2}.$$

Let $C = \{(x, y, 0) \mid x \geq 0, y \geq 0\}$. Let L be the subspace $\{(0, 0, z) \mid z \in \mathbb{R}\}$. Let $S = \{(1, 1, 1); (0, 0, 2)\} := \{\bar{x}, x'\}$. Let $\hat{x} = (0, 0, 0)$. Then, $\bar{x} \in E(S, C)$, but

$$d(\bar{x}, \hat{x} + L) = \sqrt{2} > 0 = d(x', \hat{x} + L).$$

A similar proof gives us the following result for weakly efficient points.

THEOREM 5.3. *Let X , C and L satisfy the assumptions of Theorem 5.1. Then the following assertions hold.*

(a) *Each weakly efficient point \bar{x} of an arbitrary set S fulfills the following condition: for a given \hat{x} such that $p(\hat{x}) \in p(\bar{x}) - \text{cor } C^+$,*

$$d(\bar{x}, \hat{x} + L) \leq d(x, \hat{x} + L) \quad \text{for all } x \in S$$

if and only if the norm considered on L^+ is the Minkowski functional corresponding to the order interval $[p(\hat{x}) - p(\bar{x}), p(\bar{x}) - p(\hat{x})]$.

(b) Each weakly efficient point of an arbitrary set S satisfies the following condition: for a given \tilde{x} such that $p(\tilde{x}) \in p(\bar{x}) + \text{cor } C^+$,

$$d(\bar{x}, \tilde{x} + L) \geq d(x, \tilde{x} + L) \quad \text{for all } x \in (\tilde{x} - C) \cap S$$

if and only if the norm considered on L^+ is the Minkowski functional corresponding to the order interval $[p(\bar{x}) - p(\tilde{x}), p(\tilde{x}) - p(\bar{x})]$.

Theorem 5.3(a) extends Theorem 3.2 of [7].

Remark 5.4. Using norm scalarization to obtain for proper efficiency, necessary conditions similar to Theorems 5.1 and 5.3 seem to be so complicated that they have no practical meaning. Using a penalty scalarizing functional, which is closely related to the norm, is much more convenient (see [18]).

6. Application: A control approximation problem. We investigate the following control approximation problem, which is similar to the control approximation problem considered in [10], but with a nonpointed ordering cone.

Let a control system

$$(20) \quad \dot{x}(t) = f(x(t), u(t)) \quad \text{almost everywhere on } [t_0, t_1],$$

$$(21) \quad x(t_0) = x_0,$$

$$(22) \quad u(\cdot) \in U$$

be given; where $F: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is a given mapping; $x_0 \in \mathbb{R}^n$ and $[t_0, t_1] \subset \mathbb{R}$ are given; and $U \subset L_\infty^m[t_0, t_1]$ are also given. Assume that for each control $u(\cdot) \in U$, the system (20)–(22) has a unique solution $x(\cdot, u) \in W_{1\infty}^n[t_0, t_1] := \{x: [t_0, t_1] \rightarrow \mathbb{R}^n \mid x \text{ is absolutely continuous and } \dot{x} \in L_\infty^n[t_0, t_1]\}$.

Let $z(\cdot) \in C^n[t_0, t_1]$ be given, where $C^n[t_0, t_1]$ is equipped with the norm

$$\|x(\cdot)\| = \max_{\substack{i \leq i \leq n \\ t_0 \leq t \leq t_1}} |x^i(t)|,$$

and ordered by the cone

$$C = \{x \in C^n[t_0, t_1] \mid x_i(t) \geq 0 \text{ for } i = 2, \dots, n \text{ and } t \in [t_0, t_1]\}.$$

We recall that a supremal point of a subset A of a linear ordered space X , denoted by $\sup A$, is a point $\bar{a} \in X$ such that $a \leq \bar{a}$ for all $a \in A$ and $\bar{a} \leq b$ for all b satisfying $a \leq b$ for all $a \in A$. An infimal point is similarly defined. If each pair of points of X has a supremal and an infimal point, X is called a linear lattice. A point $|x|^{or} := \sup \{x, -x\}$ is called an order absolute value of x . It is known that $C^n[t_0, t_1]$ ordered by the above C is a linear lattice.

Following [10], we say that $\bar{u}(\cdot) \in U$ is an optimal control (or weakly optimal control) if $|x(\cdot, \bar{u}) - z(\cdot)|^{or}$ is an efficient point (or weakly efficient point, respectively) of the set

$$S := \{|x(\cdot, u) - z(\cdot)|^{or} \mid u \in U\}.$$

The problem of finding an optimal control or a weakly optimal control is called a control approximation problem. Clearly, this problem corresponds to the vector optimization problem of finding an efficient point or a weakly efficient point of S .

In this case, we set $X = C^n[t_0, t_1]$,

$$L = C \cap (-C) = \{x \in X \mid x = (x^1, 0, \dots, 0)\},$$

$$L^+ = \{x \in X \mid x = (0, x^2, \dots, x^n)\}.$$

Then L^+ is orthogonal to L . Moreover $S \subseteq C$ and both (3) and (6) are satisfied. Observe also that, for $x \in C^n[t_0, t_1]$,

$$d(x, L) = \|p(x)\| = \|(O, x^2, \dots, x^n)\|,$$

$$\| |p(x)|^{or} \| = \max_{\substack{2 \leq i \leq n \\ t_0 \leq t \leq t_1}} |x^i(t)| = \|p(x)\|.$$

By virtue of Theorems 3.1 and 4.1 we obtain the following result. The proof is a direct calculation and then omitted.

COROLLARY 6.1. (a) A control $\bar{u} \in U$ is optimal if

$$\max_{\substack{2 \leq i \leq n \\ t_0 \leq t \leq t_1}} |x^i(t, \bar{u}) - z^i(t)| < \max_{\substack{2 \leq i \leq n \\ t_0 \leq t \leq t_1}} |x^i(t, u) - z^i(t)|$$

for all $u \in U$ with $|x(\cdot, u) - z(\cdot)|^{or} \neq |x(\cdot, \bar{u}) - z(\cdot)|^{or}$.

(b) A control $\bar{u} \in U$ is optimal if

$$\max_{\substack{2 \leq i \leq n \\ t_0 \leq t \leq t_1}} |x^i(t, \bar{u}) - z^i(t)| \leq \max_{\substack{2 \leq i \leq n \\ t_0 \leq t \leq t_1}} |x^i(t, u) - z^i(t)|$$

for all $u \in U$.

Acknowledgments. I would like to thank the referees and Professor J. P. Penot for many valuable remarks and suggestions.

REFERENCES

- [1] M. P. BENSON, *An improved definition of proper efficiency for vector maximization with respect to cones*, J. Math. Anal. Appl., 71 (1979), pp. 232-241.
- [2] J. BORWEIN, *Proper efficient points for maximizations with respect to cones*, SIAM J. Control Optim., 15 (1977), pp. 57-63.
- [3] V. J. BOWMAN, *On the relationship of the Tchebycheff norm and the efficient frontier of multiple-criteria objectives*, in Multiple Criteria Decision Making, H. Thiriez and S. Zionts, eds., Springer-Verlag, Berlin, 1976, pp. 76-85.
- [4] A. M. GEOFFRION, *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl., 22 (1968), pp. 618-630.
- [5] M. I. HENIG, *Proper efficiency with respect to cones*, J. Optim. Theory Appl., 36 (1982), pp. 387-407.
- [6] L. HURWICZ, *Programming in linear spaces*, in Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford Univ. Press, Stanford, 1958, pp. 38-102.
- [7] J. JAHN, *Scalarization in vector optimization*, Math. Programming, 29 (1984), pp. 203-218.
- [8] ———, *A characterization of properly minimal elements of a set*, SIAM J. Control Optim., 23 (1985), pp. 649-656.
- [9] ———, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Peter Lang, Frankfurt-am-Main, Germany, 1986.
- [10] ———, *Parametric approximation problems arising in vector optimization*, J. Optim. Theory Appl., 54 (1987), pp. 503-516.
- [11] P. Q. KHANH, *On proper solutions of vector optimization problems*, J. Optim. Theory Appl., to appear.
- [12] W. KRABS, *Optimization and Approximation*, John Wiley, Chichester, 1979.
- [13] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in Proc. Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., Univ. California Press, Berkeley, 1951, pp. 481-492.
- [14] S. ROLEWICZ, *On a norm scalarization in infinite dimensional Banach spaces*, Control Cyber., 4 (1975), pp. 85-89.
- [15] M. E. SALUKVADZE, *Vector-Valued Optimization Problems in Control Theory*, Academic Press, New York, 1979.
- [16] Y. SAWARAGI, H. NAKAYAMA, AND T. TANINO, *Theory of Multiobjective Optimization*, Academic Press, New York, 1985.

- [17] A. WIERZBICKI, *Penalty methods in solving optimization problems with vector performance criteria*, Technical Report 12, Inst. Automatic Control, Tech. Univ. Warsaw, 1974.
- [18] ———, *Basic properties of scalarizing functionals for multiobjective optimization*, Math. Operationsforsch. Statist. Ser. Optimization, 8 (1977), pp. 55–60.
- [19] ———, *On the completeness and constructiveness of parametric characterizations to vector optimization problems*, OR Spektrum, 8 (1986), pp. 73–87.
- [20] ———, *Multiple criteria solutions in noncooperative game theory, Part III: Theoretical foundations*, to appear.
- [21] P. L. YU AND G. LEITMANN, *Compromise solutions, domination structures and Savlukadze's solution*, J. Optim. Theory Appl., 13 (1974), pp. 362–378.

CONTROLLABILITY AND STABILIZABILITY OF THE THIRD-ORDER LINEAR DISPERSION EQUATION ON A PERIODIC DOMAIN*

D. L. RUSSELL† AND B. Y. ZHANG‡

Abstract. In this paper, solutions of the third-order linear dispersion equations

$$\frac{\partial w}{\partial t} + \frac{\partial^3 w}{\partial x^3} = f(x, t) \quad \text{and} \quad \frac{\partial w}{\partial t} + \frac{\partial^3 w}{\partial x^3} = 0$$

are studied for $t \geq 0, 0 \leq x \leq 2\pi$. In the first case, periodic boundary conditions are imposed at 0 and 2π and the distributed control f , which may, however, have support smaller than $[0, 2\pi]$, is assumed to be generated by a linear feedback law conserving the "volume" $\int_0^{2\pi} w(x, t) dx$ while monotonically reducing $\int_0^{2\pi} w(x, t)^2 dx$. For the second equation, a feedback boundary control having the same properties is applied. In both cases, uniform exponential decay to a constant state is obtained. Related exact controllability questions are also studied, and affirmative results are obtained.

Key words. control, stabilization, dispersion, periodic

AMS(MOS) subject classifications. 93D15, 93C20, 93B05, 35Q20

1. Introduction. In [8], [11], and in the doctoral thesis [16] of Zhang a certain control problem has been introduced for the forced Korteweg-de Vries equation

$$(1.1) \quad \frac{\partial w}{\partial t} + w \frac{\partial w}{\partial x} + \frac{\partial^3 w}{\partial x^3} = f(x, t)$$

on the domain $t \geq 0, 0 \leq x \leq 2\pi$, with periodic boundary conditions

$$(1.2) \quad \frac{\partial^k w}{\partial x^k}(0, t) = \frac{\partial^k w}{\partial x^k}(2\pi, t), \quad k = 0, 1, 2,$$

so that the process is periodic with period 2π in the variable x . Issues arising in [11] and [16] have made it clear that further progress requires a very thorough knowledge of the control theory for the related linear third-order equation

$$(1.3) \quad \frac{\partial w}{\partial t} + \frac{\partial^3 w}{\partial x^3} = f(x, t)$$

on the same domain with the same boundary conditions. Such a study is the theme of the present paper.

Let A denote the operator

$$(1.4) \quad Aw = -w'''$$

on the domain $\mathcal{D}(A) \subset L^2(0, 2\pi)$ consisting of functions in $H^3(0, 2\pi)$ satisfying boundary conditions of the form (1.2). It is immediately clear that A generates a strongly continuous unitary group on $L^2(0, 2\pi)$; the eigenfunctions are simply the orthogonal Fourier basis functions in $L^2(0, 2\pi)$,

$$\varphi_k(x) = (2\pi)^{-1/2} e^{ikx}, \quad k = 0, \pm 1, \pm 2, \dots,$$

and the corresponding exponential solutions take the form

$$(1.5) \quad w_k(x, t) = e^{i(k^3 t + kx)}$$

* Received by the editors June 3, 1991; accepted for publication (in revised form) December 18, 1991. This research was supported in part by U.S. Air Force Office of Scientific Research grant AFOSR 89-0031.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

‡ Department of Mathematics, University of Cincinnati, Cincinnati, Ohio 45221.

for the same values of k . The formula (1.5) shows immediately that the speed of propagation of sinusoidal waveforms is proportional to the cube of the spatial frequency (inversely proportional to the cube of the wavelength), resulting in rapid dispersion properties of solutions. For $f \in L^2_{\text{loc}}((0, \infty), L^2(0, 2\pi))$ and in initial state $w_0 \in L^2(0, 2\pi)$ the variation-of-parameters formula

$$(1.6) \quad w(\cdot, t) = e^{At}w_0 + \int_0^t e^{A(t-s)}f(\cdot, s) ds$$

applies to yield a strongly continuous “generalized solution” of the inhomogeneous system (1.3). As the work progresses, we will consider more specialized controls as required.

In [11] and [16] we have provided an application-oriented motivation for considering only controls that conserve the quantity

$$(1.7) \quad [w(\cdot, t)] \equiv \int_0^{2\pi} w(x, t) dx.$$

In the applications in question, this corresponds to a fluid “volume.” For appropriately smooth solutions of (1.2), (1.3), we have

$$(1.8) \quad \begin{aligned} \frac{d}{dt} \int_0^{2\pi} w(x, t) dx &= \int_0^{2\pi} \frac{\partial w}{\partial t}(x, t) dx = \int_0^{2\pi} \left(f(x, t) - \frac{\partial^3 w}{\partial x^3}(x, t) \right) dx \\ &= \int_0^{2\pi} f(x, t) dx - \frac{\partial^2 w}{\partial x^2}(x, t) \Big|_{x=0}^{x=2\pi} \end{aligned}$$

and we conclude, using the equation corresponding to $k=2$ in (1.2), that $[w(\cdot, t)]$ is conserved just in the case where

$$(1.9) \quad [f(\cdot, t)] \equiv \int_0^{2\pi} f(x, t) dx = 0$$

for almost all t . Since we are primarily interested in the linear equation (1.3) as a background for studies of the more relevant nonlinear equation (1.1), we will continue to impose this side condition on the applied controls in the linear case.

The exact controllability problem concerns the use of controls, as described, to transfer the system (1.2), (1.3) between given initial and terminal states

$$w(\cdot, 0) = w_0 \in L^2(0, 2\pi), \quad w(\cdot, T) = w_T \in L^2(0, 2\pi)$$

during the interval $[0, T]$, $T > 0$. One might also refer to this as the *open-loop* control problem.

Closed-loop control generally refers to control synthesis by state feedback of some sort and is predominantly concerned with achieving asymptotic stability of an equilibrium state or invariant set. Within the context (1.9) the appropriate invariant set is the set of constant states

$$(1.10) \quad w(x, t) \equiv c,$$

where $c = [w(\cdot, 0)]$. Since it is easy to show (see, e.g., [11]) that, among all $w \in L^2(0, 2\pi)$ such that $[w] = d$ (see (1.7) for the definition of $[w]$), for a given real d the constant state $w(x) \equiv c = d/2\pi$ has least norm in $L^2(0, 2\pi)$, we may hope to approach such a constant state c by incrementally reducing $\|w(\cdot, t)\|_{L^2(0, 2\pi)}$ with application of an appropriate control f , maintaining (1.9) so that (1.10) is true for all t . For $w(x, t)$ an

appropriately smooth solution of (1.3), we easily obtain, with two applications of integration by parts,

$$\begin{aligned}
 \frac{d}{dt} \left(\frac{1}{2} \int_0^{2\pi} w(x, t)^2 dx \right) &= \int_0^{2\pi} w(x, t) \left(f(x, t) - \frac{\partial^3 w}{\partial x^3}(x, t) \right) dx \\
 (1.11) \qquad &= \int_0^{2\pi} w(x, t) f(x, t) dx \\
 &\quad + \left(\frac{1}{2} \left(\frac{\partial w}{\partial x}(x, t) \right)^2 - w(x, t) \frac{\partial^2 w}{\partial x^2}(x, t) \right) \Big|_{x=0}^{x=2\pi}.
 \end{aligned}$$

If periodic boundary conditions (1.2) apply, we have

$$(1.12) \qquad \frac{d}{dt} \left(\frac{1}{2} \int_0^{2\pi} w(x, t)^2 dx \right) = \int_0^{2\pi} w(x, t) f(x, t) dx.$$

With $\mathcal{H} > 0$ and

$$(1.13) \qquad f(x, t) = -\mathcal{H} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right),$$

it is evident that (1.9) is satisfied, so that $[w(\cdot, t)]$ is constant and we may proceed further to compute

$$\begin{aligned}
 &\frac{d}{dt} \left\{ \frac{1}{2} \int_0^{2\pi} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right)^2 dx \right\} \\
 &= \frac{d}{dt} \left\{ \frac{1}{2} \int_0^{2\pi} w(x, t)^2 dx - \frac{1}{2\pi} [w(\cdot, t)]^2 \right\} \\
 (1.14) \qquad &= \int_0^{2\pi} w(x, t) \frac{\partial w}{\partial t}(x, t) dx \\
 &= -\mathcal{H} \int_0^{2\pi} w(x, t) \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right) dx \\
 &= -\mathcal{H} \int_0^{2\pi} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right)^2 dx,
 \end{aligned}$$

where we have used (1.7), (1.12), and (1.13) together with the mutual orthogonality of $[w(\cdot, t)]$ and $w(\cdot, t) - [w(\cdot, t)]/2\pi$. It follows that

$$(1.15) \qquad \left\| w(\cdot, t) - \frac{1}{2\pi} [w(\cdot, t)] \right\|_{L^2(0, 2\pi)}$$

decays to zero exponentially as t tends to infinity, i.e., $w(x, t)$ approaches the constant state $(2\pi)^{-1} [w(\cdot, t)]$ exponentially in the $L^2(0, 2\pi)$ norm.

A more interesting case is obtained if some *a priori* restrictions are imposed on the applied control $f(x, t)$. Let us suppose that $g(x)$ is a piecewise-continuous non-negative function defined for x in $[0, 2\pi]$ such that

$$(1.16) \qquad [g] = \int_0^{2\pi} g(x) dx = 1,$$

and let us restrict attention to controls of the form

$$(1.17) \qquad f(x, t) = g(x) \varphi(x, t), \quad \varphi \in L^2_{\text{loc}}((0, \infty); L^2(0, 2\pi)),$$

so that φ is, in effect, the applied control. If $[a, b]$ is a subinterval of $[0, 2\pi]$, setting $g(x) = (b - a)^{-1} \chi_{[a, b]}$ corresponds to restriction of the support of the control f .

We will study controllability of the system (1.3) by controls (1.17) in § 2. For the moment, we continue to discuss volume-conserving controls given in feedback form. With f of the form (1.13) and \mathcal{K} a positive constant, let us suppose φ to be generated by

$$(1.18) \quad \varphi(x, t) = -\mathcal{K} \left(w(x, t) - \int_0^{2\pi} g(s) w(s, t) ds \right), \quad \mathcal{K} > 0.$$

Using (1.16) we verify readily that (1.9) holds, so that $[w(\cdot, t)]$ is constant. Then, as in (1.12) and (1.14), we may compute

$$(1.19) \quad \begin{aligned} & \frac{d}{dt} \left\{ \frac{1}{2} \int_0^{2\pi} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right)^2 dx \right\} \\ &= \int_0^{2\pi} w(x, t) \frac{\partial w}{\partial t}(x, t) dx \\ &= -\mathcal{K} \int_0^{2\pi} w(x, t) g(x) \left(w(x, t) - \int_0^{2\pi} g(s) w(s, t) ds \right) dx. \end{aligned}$$

Using (1.16) we see that

$$\begin{aligned} & \int_0^{2\pi} g(x) \int_0^{2\pi} g(s) w(s, t) ds \left(w(x, t) - \int_0^{2\pi} g(s) w(s, t) ds \right) dx \\ &= \left(1 - \int_0^{2\pi} g(x) dx \right) \left(\int_0^{2\pi} g(s) w(s, t) ds \right)^2 = 0, \end{aligned}$$

so that (1.19) yields

$$(1.20) \quad \begin{aligned} & \frac{d}{dt} \left\{ \frac{1}{2} \int_0^{2\pi} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right)^2 dx \right\} \\ &= -\mathcal{K} \int_0^{2\pi} g(x) \left(w(x, t) - \int_0^{2\pi} g(s) w(s, t) ds \right)^2 dx \leq 0 \end{aligned}$$

and it is reasonable to suppose, again, that $w(\cdot, t)$ converges, in an appropriate sense, to the constant state $[w(\cdot, t)]$ as $t \rightarrow \infty$. With some further assumptions, we will establish this rigorously in § 2; we there show that with appropriate assumptions on g we have exponential decay to the constant state, uniform with respect to the norm in $L^2(0, 2\pi)$.

It is also possible to pose a volume-conserving open-loop boundary control problem for this system. Let us take $f(x, t) \equiv 0$ and apply the boundary conditions (1.2) for $k = 0$ and $k = 2$ but not for $k = 1$. Then we have

$$(1.21) \quad \frac{\partial w}{\partial t} + \frac{\partial^3 w}{\partial x^3} = 0$$

and (1.8) shows that $[w(\cdot, t)]$ is constant. Again proceeding as in (1.11) and (1.14), we have

$$(1.22) \quad \frac{d}{dt} \left\{ \int_0^{2\pi} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right)^2 dx \right\} = \left(\frac{\partial w}{\partial x}(x, t) \right)^2 \Big|_{x=0}^{x=2\pi}.$$

If we suppose that the boundary control mechanism takes the form

$$(1.23) \quad \frac{\partial w}{\partial x}(2\pi, t) - \frac{\partial w}{\partial x}(0, t) = h(t),$$

which, since 2π and 0 are identified in this context, is the same as

$$\frac{\partial w}{\partial x}(0^-, t) - \frac{\partial w}{\partial x}(0^+, t) = h(t),$$

then (1.21) becomes

$$\frac{d}{dt} \left\{ \int_0^{2\pi} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right)^2 dx \right\} = \left(\frac{\partial w}{\partial x}(2\pi, t) + \frac{\partial w}{\partial x}(0, t) \right) h(t).$$

An appropriate feedback mechanism is then

$$(1.24) \quad h(t) = -\mathcal{K} \left(\frac{\partial w}{\partial x}(2\pi, t) + \frac{\partial w}{\partial x}(0, t) \right), \quad \mathcal{K} > 0,$$

resulting in

$$(1.25) \quad \frac{d}{dt} \left\{ \frac{1}{2} \int_0^{2\pi} \left(w(x, t) - \frac{1}{2\pi} [w(\cdot, t)] \right)^2 dx \right\} = -\mathcal{K} \left(\frac{\partial w}{\partial x}(2\pi, t) + \frac{\partial w}{\partial x}(0, t) \right)^2 \leq 0.$$

Again we suspect that $w(\cdot, t)$ tends to the constant state $[w(\cdot, t)]$ as $t \rightarrow \infty$; we study this question in § 3 and show that, in fact, such decay does hold and is uniformly exponential with respect to the norm in $L^2(0, 2\pi)$. Concluding that section, we note that the uniform exponential decay to the constant state obtained for (1.23) and (1.24) implies exact controllability between initial and terminal states w_0 and $w_T \in L^2(0, 2\pi)$ for T appropriately large but independent of these states, provided only that $[w_0] = [w_T]$.

It is not difficult to see that the control mechanism (1.23) corresponds to the application of a dipole of strength $h(t)$ at the point 0 or, equivalently, 2π . One may proceed further to show that control processes of the form (1.17) can be regarded as spatial convolutions of the process (1.23) by using appropriate convolution kernels. We will not pursue these matters further here.

2. Exponential decay rates with distributed controls of restricted form. If we define the operator A as in (1.4), then the system (1.3) with $f(x, t)$ given by (1.17) and (1.18) can be expressed in the form

$$(2.1) \quad \dot{w} = (A - \mathcal{K}G)w$$

(where the dot signifies d/dt), with

$$(2.2) \quad (Gw)(x) = g(x) \left(w(x) - \int_0^{2\pi} g(s)w(s) ds \right), \quad w \in L^2(0, 2\pi).$$

Since A generates a strongly continuous semigroup $S(t)$ of bounded operators in $L^2(0, 2\pi)$ and G is a bounded \cdot operator, $A - \mathcal{K}G$ also generates a strongly continuous

semigroup $S_{\mathcal{K}}(t)$ on that space (see [6]). Thus, given an initial state $w_0 \in L^2(0, 2\pi)$, we have a unique solution $w(t)$, continuous with respect to the norm in that space for $t \geq 0$, given by

$$(2.3) \quad w(t) = S_{\mathcal{K}}(t)w_0.$$

Further results from the standard semigroup theory show that if $w_0 \in H_p^3(0, 2\pi)$, the domain of A and hence of $A - \mathcal{K}G$, the resulting solution w lies in $C((0, \infty); H_p^3(0, 2\pi)) \cap C^1((0, \infty); L^2(0, 2\pi))$.

From the work of § 1 we know that $[w(\cdot, t)]$ is constant and therefore equal to $[w_0]$ for all $t \geq 0$. Then (1.19) leads us to suspect that $w(\cdot, t)$ converges to the constant state $[w_0] \equiv [w(\cdot, t)]$ as $t \rightarrow \infty$. We will show that this is the case and that, in fact, the decay to the constant state takes place at a uniform exponential rate.

From the general plan outlined in [12], [14], [15], the first step in obtaining exponential decay for the system (1.3), (1.17), (1.18) is to obtain an exact controllability result for the related system

$$(2.4) \quad \frac{\partial v}{\partial t} + \frac{\partial^3 v}{\partial x^3} = Gf,$$

with v subject to the boundary conditions (1.2) and with the operator G as in (2.2). We may state this result as the following theorem.

THEOREM 1. *Let $T > 0$ be given, and let it be assumed that the function g associated with G is of class C^0 on $[0, 2\pi]$ (is of class $C_p^3(g, g', \text{ and } g'' \text{ periodic})$ on $[0, 2\pi]$). Given any final state $v_T \in L^2(0, 2\pi)$ ($v_1 \in H_p^3(0, 2\pi)$) with (see (1.7)) $[v_T] = 0$, there exists a control $f \in L^2((0, T); L^2(0, 2\pi))$ ($f \in L^2((0, T); H_p^3(0, 2\pi))$) such that the solution v of (1.2) and (2.4) with*

$$(2.5) \quad v(\cdot, 0) = 0 \quad \text{in } L^2(0, 2\pi) \ (H_p^3(0, 2\pi))$$

satisfies the terminal condition

$$(2.6) \quad v(\cdot, T) = v_T \quad \text{in } L^2(0, 2\pi) \ (H_p^3(0, 2\pi)).$$

Moreover, there is a positive number C , independent of v_T , such that

$$\|v\|_{L^2((0, T); L^2(0, 2\pi))} \leq C \|v_T\|_{L^2(0, 2\pi)} \quad (\|v\|_{L^2((0, T); H_p^3(0, 2\pi))} \leq C \|v_T\|_{H_p^3(0, 2\pi)}).$$

Proof. It is well known that the operator A as defined in (1.4) has eigenvalues $\lambda_k = -ik^3$ corresponding to eigenfunctions

$$\varphi_k(x) = e^{ikx}, \quad -\infty < k < \infty.$$

Relative to this basis, the terminal state v_T has the expansion, convergent in $L^2(0, 2\pi)$,

$$(2.7) \quad v_T = \sum_{k=-\infty}^{\infty} v_k \varphi_k, \quad v_k = \frac{1}{2\pi} \int_0^{2\pi} v_T(x) \overline{\varphi_k(x)} dx.$$

The homogeneous equation

$$(2.8) \quad \frac{\partial u}{\partial t} + \frac{\partial^3 u}{\partial x^3} = 0,$$

with periodic boundary conditions (1.2) has corresponding solutions

$$(2.9) \quad u_k(\cdot, t) = e^{\lambda_k t} \varphi_k.$$

For smooth f , periodic on $[0, 2\pi]$, we readily compute, using integration by parts with v satisfying (2.4) and (1.2), that

$$\frac{d}{dt} \int_0^{2\pi} v(x, t) \overline{u_k(x, t)} dx = \int_0^{2\pi} (Gf)(x, t) \overline{u_k(x, t)} dx.$$

Integrating with respect to t , we have

$$(2.10) \quad \begin{aligned} & \int_0^{2\pi} v(x, T) \overline{u_k(x, T)} dx - \int_0^{2\pi} v(x, 0) \overline{u_k(x, 0)} dx \\ &= \int_0^T \int_0^{2\pi} (Gf)(x, t) \overline{u_k(x, t)} dx dt. \end{aligned}$$

Continuity considerations then show that (2.10) continues to be valid for $f \in L^2((0, T); L^2(0, 2\pi))$.

Evaluation of the integrals in (2.10) with

$$\tilde{v}_k = \frac{1}{2\pi} \int_0^{2\pi} v(x, T) \overline{\varphi_k(x)} dx$$

and u_k as in (2.9) shows that

$$(2.11) \quad 2\pi \tilde{v}_k = \int_0^T e^{\lambda_k(T-t)} \int_0^{2\pi} (Gf)(x, t) \overline{\varphi_k(x)} dx dt, \quad -\infty < k < \infty.$$

By defining $p_k(t) = e^{\lambda_k t}$, $\mathcal{P} \equiv \{p_k \mid -\infty < k < \infty\}$ may be seen, from the results in [5], to form a Riesz basis for its closed span, P_T , in $L^2(0, T)$. We let $\mathcal{Q} \equiv \{q_k \mid -\infty < k < \infty\}$ be the unique dual Riesz basis for \mathcal{P} in P_T , i.e., the functions in \mathcal{Q} are the unique elements of P_T such that

$$(2.12) \quad \int_0^T q_j(t) \overline{p_k(t)} dt = \delta_{kj}, \quad -\infty < j, k < \infty.$$

We take the control f in (2.4) to have the form

$$(2.13) \quad f(x, t) = \sum_{j=-\infty}^{\infty} f_j q_j(t) (G\varphi_j)(x),$$

where the coefficients f_j are to be determined so that, among other things, the series (2.13) is appropriately convergent. Substituting (2.13) into (2.11) yields, by using the biorthogonality (2.12),

$$(2.14) \quad \begin{aligned} 2\pi \tilde{v}_k &= e^{\lambda_k T} \sum_{j=-\infty}^{\infty} f_j \int_0^T \overline{e^{\lambda_k t}} q_j(t) \int_0^{2\pi} G(G\varphi_j)(x) \overline{\varphi_k(x)} dx dt \\ &= f_k e^{\lambda_k T} \int_0^{2\pi} G(G\varphi_k)(x) \overline{\varphi_k(x)} dx, \quad -\infty < k < \infty. \end{aligned}$$

We verify easily that G is a selfadjoint operator on $L^2(0, 2\pi)$, so that

$$\int_0^{2\pi} G(G\varphi_k)(x) \overline{\varphi_k(x)} dx = (\|G\varphi_k\|_{L^2(0, 2\pi)})^2, \quad -\infty < k < \infty.$$

Since $|\varphi_k(x)| \equiv 1$, for $-\infty < k < \infty$ we have

$$\begin{aligned} (\|G\varphi_k\|_{L^2(0, 2\pi)})^2 &= \int_0^{2\pi} \left| g(x) \left(\varphi_k(x) - \int_0^{2\pi} g(s) \varphi_k(s) ds \right) \right|^2 dx \\ &= \int_0^{2\pi} g(x)^2 dx - 2 \left| \int_0^{2\pi} g(x) \varphi_k(x) dx \right|^2 \end{aligned}$$

$$+ \int_0^{2\pi} g(x)^2 dx \left| \int_0^{2\pi} g(x) \varphi_k(x) dx \right|^2 \equiv g_k.$$

Since $\varphi_0(x) \equiv 1$ it is easy to see, by using (1.16), that $g_0 = 0$. For $k \neq 0$ the fact that $g(x)$ is real valued shows that $g(x)\varphi_k(x)$ cannot be constant on any interval, so that $g_k \neq 0$. The familiar Lebesgue lemma shows that

$$(2.15) \quad \lim_{k \rightarrow \infty} g_k = \int_0^{2\pi} g(x)^2 dx \neq 0.$$

Clearly, \tilde{v}_0 must be zero since $g_0 = 0$. From (2.15) and the fact that $g_k \neq 0$ for $k \neq 0$, it follows that there is a positive δ such that $|g_k| > \delta^2$ for $k \neq 0$. If we set $f_0 = 0$ and

$$(2.16) \quad f_k = \frac{2\pi e^{-\lambda_k T} v_k}{g_k}, \quad -\infty < k < \infty, \quad k \neq 0.$$

Equation (2.14) implies $\tilde{v}_k = v_k$, where v_k is given by (2.7). Since the v_k are square summable and the g_k are bounded below, the f_k are also square summable and (2.16) gives

$$\begin{aligned} (\|f\|_{L^2((0,T);L^2(0,2\pi))})^2 &= \int_0^T \int_0^{2\pi} |f(x,t)|^2 dx dt \\ &= \int_0^{2\pi} \int_0^T \left| \sum_{k=-\infty}^{\infty} f_k q_k(t) (G\varphi_k)(x) \right|^2 dt dx \\ &\leq 2\pi\gamma^2 \int_0^T \left| \sum_{k=-\infty}^{\infty} f_k q_k(t) \right|^2 dt \\ (2.17) \quad &\leq 2\pi\gamma^2 C^2 \sum_{k=-\infty}^{\infty} |f_k|^2 \\ &\leq (2\pi)^3 \frac{\gamma^2}{\delta^2} C^2 |v_k|^2 \\ &= 4\pi^2 \frac{\gamma^2}{\delta^2} C^2 (\|v_T\|_{L^2(0,2\pi)})^2, \end{aligned}$$

where the constant C comes from the Riesz basis property of \mathcal{Q} in P_T , δ^2 is the lower bound for the g_k referred to previously, and

$$\gamma = \sup_{\substack{x \in [0, 2\pi] \\ -\infty < k < \infty}} |(G\varphi_k)(x)| < \infty.$$

A similar computation using only terms corresponding to $|k| > K$ from the series (2.13) shows that the series converges in $L^2((0, T); (0, 2\pi))$. We conclude that f , represented by (2.13), is an element of the space $L^2((0, T); L^2(0, 2\pi))$ bounded in terms of v_T as indicated in (2.17).

With A as defined in (1.4), the solution $v(\cdot, t)$ corresponding to the control $f(\cdot, t)$ just constructed has the form

$$(2.18) \quad v(\cdot, t) = \int_0^t e^{-A(t-s)} Gf(\cdot, s) ds.$$

Since G is a bounded operator and for $t \in [0, T]$

$$\|f\|_{L^1((0,t);L^2(0,2\pi))} \leq t^{1/2} \|f\|_{L^2((0,T);L^2(0,2\pi))},$$

we conclude that (2.18) converges for each such t and, using (2.17), we can see that there is a positive number B such that

$$(2.19) \quad \|v(\cdot, t)\|_{L^2(0, 2\pi)} \leq B \|v_T\|_{L^2(0, 2\pi)}, \quad t \in [0, T].$$

An argument similar to that found in [3] shows that $v(\cdot, t)$ is continuous, as a function of t , relative to the norm in $L^2(0, 2\pi)$. The identity (2.10) with the indicated u_k and our choice of f shows that $v(\cdot, T) = v_T$.

Now suppose $g(x)$ has a piecewise-continuous third derivative and that $v_T \in H_p^3(0, 2\pi)$. The coefficients v_k in (2.7) will have the form

$$v_k = (|k| + 1)^{-3} \hat{v}_k,$$

where the \hat{v}_k are square summable. Then, from (2.16) we have

$$f_k = (|k| + 1)^{-3} \hat{f}_k,$$

where the \hat{f}_k are square summable. Since $g(x)$ has a piecewise-continuous third derivative, it is a simple matter to verify, just as for the original series (2.13), that the series

$$\frac{\partial^3 f}{\partial x^3}(x, t) = \sum_{k=-\infty}^{\infty} \hat{f}_k q_k(t) (|k| + 1)^{-3} (G\varphi_k)'''(x)$$

converges in $L^2((0, T); L^2(0, 2\pi))$. Noting that (2.4) can be written

$$\frac{\partial v}{\partial t} + \frac{\partial^3 v}{\partial x^3} = g(x) \left(f(x, t) - \int_0^{2\pi} g(\xi) f(\xi, t) d\xi \right)$$

and differentiating three times with respect to x , we obtain an equation of the form

$$\frac{\partial}{\partial t} \left(\frac{\partial^3 v}{\partial x^3} \right) + \frac{\partial^3}{\partial x^3} \left(\frac{\partial^3 v}{\partial x^3} \right) = \hat{f}(x, t),$$

with \hat{f} in $L^2((0, T); L^2(0, 2\pi))$, which, together with the periodic boundary conditions and the initial state

$$\frac{\partial^3 v}{\partial x^3}(x, 0) \equiv 0,$$

allows us to replace the estimate (2.19) by an estimate

$$(2.20) \quad \|v(\cdot, t)\|_{H_p^3(0, 2\pi)} \leq B \|v_T\|_{H_p^3(0, 2\pi)}, \quad t \in [0, T].$$

With this the proof of Theorem 1 is complete. \square

We are now in a position to obtain an exponential decay result for the system (2.1), (2.2).

THEOREM 2. *There exist $\alpha, \beta > 0$ such that, for any $w_0 \in L^2(0, 2\pi)$ the unique solution $w(\cdot, t)$ of (1.3), (1.17), (1.18), equivalently (2.1), (2.2), satisfies*

$$(2.21) \quad \|w(\cdot, t) - [w_0]\|_{L^2(0, 2\pi)} \leq \beta e^{-\alpha t} \|w_0 - [w_0]\|_{L^2(0, 2\pi)}, \quad t \geq 0.$$

Proof. We assume without loss of generality that since $w(\cdot, t) - [w_0]$ is a solution of our system, then $[w_0] = 0$. Furthermore, we consider only real solutions here and dispense with the conjugate notation used in the inner products of the preceding theorem, where complex solutions were necessarily discussed.

Let $T > 0$, and let $w(\cdot, t)$ be as described. Then $w(\cdot, T) \in L^2(0, 2\pi)$, and if $w_0 \in H_p^3(0, 2\pi) = \mathcal{D}(A_G)$, then $w(\cdot, t) \in H_p^3(0, 2\pi)$ for all $t \geq 0$, in particular, for $t = T$. From Theorem 1 we see that we can find $f(\cdot, t) \in L^2(0, 2\pi)$, $t \in [0, T]$, such that the resulting solution $v(\cdot, t)$ of (2.4), (2.5), (1.2) satisfies

$$v(\cdot, T) = w(\cdot, T)$$

and from (2.17) and (2.19) there are positive numbers B and D such that

$$(2.22) \quad \|f\|_{L^2((0,T);L^2(0,2\pi))} \leq D \|w(\cdot, T)\|_{L^2(0,2\pi)}, \quad t \in [0, T],$$

$$(2.23) \quad \|v(\cdot, t)\|_{L^2(0,2\pi)} \leq B \|w(\cdot, T)\|_{L^2(0,2\pi)}, \quad t \in [0, T].$$

Furthermore, if w_0 , and hence $w(\cdot, T)$, lies in $H_p^3(0, 2\pi)$, then (2.22) and (2.23) remain true with $L^2(0, 2\pi)$ replaced by $H_p^3(0, 2\pi)$ where ever it occurs. As we have seen in § 1 the computations carried out there are immediately valid for $w_0 \in H_p^3(0, 2\pi)$ and follow from continuity considerations if $w_0 \in L^2(0, 2\pi)$,

$$(2.24) \quad \begin{aligned} & \int_0^{2\pi} w(x, T)^2 dx - \int_0^{2\pi} w_0(x)^2 dx \\ &= -\mathcal{H} \int_0^T \int_0^{2\pi} g(x) \left(w(x, t) - \int_0^{2\pi} g(s) w(s, t) ds \right)^2 dx dt. \end{aligned}$$

An analogous computation using (1.3), (1.2), and (2.4) yields

$$(2.25) \quad \begin{aligned} \int_0^{2\pi} w(x, T)^2 dx &= \int_0^{2\pi} w(x, T) v(x, T) dx - \int_0^{2\pi} w(x, 0) v(x, 0) dx \\ &= \int_0^T \int_0^{2\pi} w(x, t) (Gf)(x, t) dx dt - \mathcal{H} \int_0^T \int_0^{2\pi} (Gw)(x, t) v(x, t) dx dt \\ &= \int_0^T \int_0^{2\pi} (Gw)(x, t) (f(x, t) - \mathcal{H}v(x, t)) dx dt \\ &\leq \|Gw\|_{L^2((0,T);(0,2\pi))} \|f - \mathcal{H}v\|_{L^2((0,T);(0,2\pi))} \\ &\leq \left[\hat{g} \int_0^T \int_0^{2\pi} g(x) \left(w(x, t) - \int_0^{2\pi} g(s) w(s, t) ds \right)^2 dx dt \right]^{1/2} \\ &\quad \cdot \|f - \mathcal{H}v\|_{L^2((0,T);(0,2\pi))}, \end{aligned}$$

where $\hat{g} > 0$ is the least upper bound for $g(x)$ in the interval $[0, 2\pi]$. Combining (2.22) with (2.23), we can see that there is a positive number E such that

$$(2.26) \quad \|f - \mathcal{H}v\|_{L^2((0,T);(0,2\pi))}^2 \leq E \|w(\cdot, T)\|_{L^2(0,2\pi)}^2.$$

Substituting (2.26) into (2.25), we have

$$\int_0^T \int_0^{2\pi} g(x) \left(w(x, t) - \int_0^{2\pi} g(s) w(s, t) ds \right)^2 dx dt \leq (E\hat{g})^{-1} \|w(\cdot, T)\|_{L^2(0,2\pi)}^2.$$

Substituting this, in turn, into (2.24), we have

$$\int_0^{2\pi} w(x, T)^2 dx - \int_0^{2\pi} w_0(x)^2 dx \leq -\mathcal{H} (E\hat{g})^{-1} \|w(\cdot, T)\|_{L^2(0,2\pi)}^2,$$

so that

$$\|w(\cdot, T)\|_{L^2(0,2\pi)}^2 \leq \frac{E\hat{g}}{E\hat{g} + \mathcal{H}} \|w_0\|_{L^2(0,2\pi)}^2.$$

Repeating this estimate on successive intervals $[(k-1)T, kT]$ with w_0 and $w(\cdot, T)$ replaced by $w(\cdot, (k-1)T)$ and $w(\cdot, kT)$, respectively, and using the monotonicity of $\|w(\cdot, t)\|_{L^2(0, 2\pi)}$, we readily obtain the estimate (2.21) (recalling that we have taken, without loss of generality, $[w_0] = 0$).

Exponential decay in $H_p^3(0, 2\pi)$ for w_0 in that space is obtained quite directly; $(A - \mathcal{H}G)w_0$ lies in $L^2(0, 2\pi)$ under these circumstances, and the solution with this initial state, $(A - \mathcal{H}G)w$, by the result just established, decays exponentially in $L^2(0, 2\pi)$ to the constant state $[(A - \mathcal{H}G)w_0] = 0$. Denoting by $L_0^2(0, 2\pi)$ the orthogonal complement of the constant states in $L^2(0, 2\pi)$, we may see that $(A - \mathcal{H}G): H_p^3(0, 2\pi) \rightarrow L_0^2(0, 2\pi)$ is bounded and boundedly invertible, from which the desired result follows immediately. \square

3. Exponential decay for the equation with boundary dissipation. The question to be considered here has been developed in § 1. Thus we consider the system (1.3), (1.23), and (1.24), the last two of which result in the closed-loop boundary condition

$$(3.1) \quad \frac{\partial w}{\partial x}(2\pi, t) = \alpha \frac{\partial w}{\partial x}(0, t),$$

where $\alpha = (1 - \mathcal{H})/(1 + \mathcal{H})$ has absolute value less than unity as a consequence of the assumed positivity of \mathcal{H} .

THEOREM 3. *The operator $A_{\mathcal{H}}$ defined by $A_{\mathcal{H}}w = -w'''$, with domain consisting of functions in $H^3(0, 2\pi)$ satisfying (3.1) and (1.2) for $k=0, 2$, generates a strongly continuous semigroup $\hat{S}_{\mathcal{H}}(t)$, $t \geq 0$, of bounded operators on $L^2(0, 2\pi)$. Letting P denote the projection on $L^2(0, 2\pi)$ defined (see (1.7)) by $Pw = [w]$, we have*

$$(3.2) \quad P\hat{S}_{\mathcal{H}}(t) = \hat{S}_{\mathcal{H}}(t)P = P, \quad t \geq 0,$$

are there are positive numbers M and γ such that

$$(3.3) \quad \|\hat{S}_{\mathcal{H}}(t) - P\| \leq Me^{-\gamma t}, \quad t \geq 0.$$

Proof. The formula (1.22) shows that for $w \in \mathcal{D}(A_{\mathcal{H}})$ we have

$$(3.4) \quad (w, A_{\mathcal{H}}w) + (A_{\mathcal{H}}w, w) = |w'(x)|^2 \Big|_{x=0}^{x=2\pi} = (\alpha^2 - 1)|w'(0)|^2 \leq 0.$$

Then, for $\lambda > 0$ and $w \in \mathcal{D}(A_{\mathcal{H}})$

$$(3.5) \quad \|(\lambda I - A_{\mathcal{H}})w\|^2 = \lambda^2\|w\|^2 - \lambda((w, A_{\mathcal{H}}w) + (A_{\mathcal{H}}w, w)) + \|A_{\mathcal{H}}w\|^2 \geq \lambda^2\|w\|^2,$$

so that, with $R(\lambda, A_{\mathcal{H}})$ denoting the resolvent of $A_{\mathcal{H}}$, we have

$$\|R(\lambda, A_{\mathcal{H}})\| \leq \lambda^{-1}, \quad \lambda > 0.$$

By the Lumer-Phillips theorem [9] or the Hille-Yosida theorem [2], $A_{\mathcal{H}}$ generates a strongly continuous semigroup on $L^2(0, 2\pi)$ if the range of $\lambda I - A_{\mathcal{H}}$ is all of $L^2(0, 2\pi)$ for $\lambda > 0$. The uniform exponential decay may then be obtained from the result of Huang [4] by demonstrating the uniform boundedness of $R(\lambda, A_{\mathcal{H}})$ for λ on the imaginary axis of the complex plane. We will carry out these two tasks by explicit construction of the resolvent operator. To that end we consider the equation

$$(3.6) \quad w''' + \lambda w = f, \quad f \in L^2(0, 2\pi).$$

For $\lambda \neq 0$ we denote the three cube roots of $-\lambda$ by μ_0, μ_1 , and μ_2 ; later we will specify which root corresponds to each symbol μ_j . We define $w' = u$, $w'' = v$, and we have

$$(3.7) \quad W' \equiv \begin{pmatrix} w' \\ u' \\ v' \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\lambda & 0 & 0 \end{pmatrix} \begin{pmatrix} w \\ u \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ f \end{pmatrix} \equiv F(\lambda)W + \varphi.$$

We diagonalize the system with the transformation $(\mu = (\mu_0, \mu_1, \mu_2))$

$$(3.8) \quad W \equiv \begin{pmatrix} w \\ u \\ v \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ \mu_0 & \mu_1 & \mu_2 \\ \mu_0^2 & \mu_1^2 & \mu_2^2 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \\ \zeta \end{pmatrix} \equiv V(\mu) \Xi.$$

Then we have

$$(3.9) \quad \Xi' = D(\mu) \Xi + \psi,$$

where $D(\mu) = \text{diag}(\mu_0, \mu_1, \mu_2)$ and $\psi = V(\mu)^{-1} \varphi$, so that for $0 \leq x \leq 2\pi$

$$(3.10) \quad \Xi(x) = e^{x D(\mu)} \Xi(0) + \int_0^x e^{(x-\xi) D(\mu)} \psi(\xi) d\xi.$$

Now, with $B(\alpha) = \text{diag}(1, \alpha, 1)$ the boundary conditions take the form

$$W(2\pi) = B(\alpha) W(0); \quad \text{i.e., } \Xi(2\pi) = V(\mu)^{-1} B(\alpha) V(\mu) \Xi(0)$$

and, for $x = 2\pi$, (3.10) assumes the form

$$(3.11) \quad (B(\alpha) V(\mu) - V(\mu) e^{2\pi D(\mu)}) \Xi(0) = V(\mu) \int_0^{2\pi} e^{(2\pi-x) D(\mu)} \psi(x) dx.$$

Let

$$(3.12) \quad U(\mu, \alpha) = (B(\alpha) V(\mu) - V(\mu) e^{2\pi D(\mu)}),$$

$$(3.13) \quad \Delta(\mu, \alpha) = \det(B(\alpha) V(\mu) - V(\mu) e^{2\pi D(\mu)}).$$

If $\Delta(\mu, \alpha) = 0$, any nonzero solution $\Xi(0)$ of $U(\mu, \alpha) \Xi(0) = 0$ yields an eigenfunction

$$(3.14) \quad \Xi(\mu, \alpha, x) = e^{x D(\mu)} \Xi(0)$$

of $A_{\mathcal{H}}$ corresponding to the value of λ associated with μ by $-\lambda = \mu_j^3$, $j = 0, 1, 2$. On the other hand, if $\Delta(\mu, \alpha) \neq 0$ we can solve (3.11) for $\Xi(0)$,

$$(3.15) \quad \Xi(0) = U(\mu, \alpha)^{-1} V(\mu) \int_0^{2\pi} e^{(2\pi-s) D(\mu)} \psi(s) ds,$$

and then substitute in (3.10) to obtain

$$(3.6) \quad \begin{aligned} \Xi(\mu, \alpha, \psi, x) &= e^{x D(\mu)} U(\mu, \alpha)^{-1} V(\mu) \int_0^{2\pi} e^{(2\pi-s) D(\mu)} \psi(s) ds \\ &\quad + \int_0^x e^{(x-s) D(\mu)} \psi(s) ds. \end{aligned}$$

The corresponding linear map from f to $w(\lambda, \alpha, f, \cdot)$ defines $R(\lambda, A_{\mathcal{H}})$ as a bounded linear operator on $L^2(0, 2\pi)$ for those μ such that $\Delta(\mu, \alpha) \neq 0$. There can be no positive λ such that $\Delta(\mu, \alpha) = 0$; if that were the case, the $w(\lambda, \alpha, x)$ corresponding to $\Xi(\mu, \alpha, x)$ in (3.16) would violate (3.5). We conclude therefore that $R(\lambda, A_{\mathcal{H}})$ is defined as a bounded linear operator on $L^2(0, 2\pi)$ for $\lambda > 0$, so that $A_{\mathcal{H}}$ generates a strongly continuous semigroup on $L^2(0, 2\pi)$.

There remains the question of the uniform exponential decay result (3.3). For this, as we have indicated, we must estimate $R(\lambda, A_{\mathcal{H}})$ for λ on the imaginary axis. More accurately, as already indicated in (3.3), we want to demonstrate the exponential decay of $\hat{S}_{\mathcal{H}}(t)$ to the orthogonal projection P . Since P commutes with $A_{\mathcal{H}}$ (this is just the “volume” conservation (1.8) of § 1 with $f \equiv 0$), what is needed is a demonstration of the uniform boundedness of $R(\lambda, A_{\mathcal{H}})(I - P)$ for λ on the imaginary axis.

We argue first that the only point on the imaginary axis where $R(\lambda, A_{\mathcal{H}})$ fails to exist as a bounded linear operator on $L^2(0, 2\pi)$ is the point $\lambda = 0$. If there were a point $\lambda = i\omega$, $\omega \neq 0$, such that $R(i\omega, A_{\mathcal{H}})$ should fail to be so defined, our previous arguments show that with μ_ρ , the corresponding vector whose components are the cube roots of $-i\omega$, we should have an eigenvector $w(i\omega, \alpha, x)$ of $A_{\mathcal{H}}$ corresponding to the eigenvalue $i\omega$. It is then easy to see from (3.4) that $w'(i\omega, \alpha, 0) = w'(i\omega, \alpha, 2\pi) = 0$. But then $w(i\omega, \alpha, x)$ is also an eigenfunction of A_0 , the third-order operator with periodic boundary conditions, corresponding to the eigenvalue $i\omega$. Since all such eigenfunctions are known to be nontrivial exponential functions unless $\omega = 0$, $w'(i\omega, \alpha, 0) = 0$ is impossible unless $\omega = 0$, and we see that the resolvent $R(\lambda, A_{\mathcal{H}})$ is indeed defined as a bounded linear operator on $L^2(0, 2\pi)$ for $\lambda \neq 0$ on the imaginary axis.

That being the case, to complete the proof of (3.3) by means of Huang's cited result, it is sufficient to show (i) that $R(\lambda, A_{\mathcal{H}})$ is uniformly bounded for large λ on the imaginary axis and (ii) that $R(\lambda, A_{\mathcal{H}})(I - P)$ remains bounded as $\lambda \rightarrow 0$. We will prove (i)—indeed, we will prove an even stronger result—in Lemma 4 to follow. In anticipation of that result, the proof of Theorem 3 requires only the demonstration of (ii).

The Hilbert space $L^2(0, 2\pi)$ has the orthogonal decomposition

$$L^2(0, 2\pi) = L_0^2(0, 2\pi) + [L^2](0, 2\pi),$$

where $L_0^2(0, 2\pi)$ denotes those $f \in L^2(0, 2\pi)$ for which $[f] = 0$ and $[L^2]$ is the subspace of constant functions in $L^2(0, 2\pi)$. Then P (see text preceding (3.2)) and $Q = I - P$ are the orthogonal projections on these subspaces. The space $[L^2]$, interpreted as a subspace of $H_p^3(0, 2\pi)$, is the null space of $A_{\mathcal{H}}$, and $L_0^2(0, 2\pi) \cap H_p^3(0, 2\pi)$ is a closed invariant subspace for $A_{\mathcal{H}}$ (this is implied by the computation (1.8)). The identities (3.2) follow directly from these observations, and (3.3) follows if we can show that the semigroup $\tilde{S}_{\mathcal{H}}(t)$ generated on $L^2(0, 2\pi)$ by the restriction $\tilde{A}_{\mathcal{H}}$ of $A_{\mathcal{H}}$ to $L_0^2(0, 2\pi)$ decays exponentially. The bounds on $R(\lambda, A_{\mathcal{H}})$ for λ on the imaginary axis bounded away from zero (to be proved in Lemma 4, below) continue to apply to $R(\lambda, \tilde{A}_{\mathcal{H}})$ there. To apply the cited theorem of Huang to $\tilde{S}_{\mathcal{H}}(t)$, it is therefore only necessary to show that we can obtain a bound

$$(3.17) \quad \|R(\lambda, \tilde{A}_{\mathcal{H}})f\|_{L_0^2(0, 2\pi)} = \|R(\lambda, A_{\mathcal{H}})f\| \leq K \|f\|_{L^2(0, 2\pi)}, \quad f \in L_0^2(0, 2\pi),$$

for some positive K independent of f , for λ near zero.

From (3.7) and (3.16), we have

$$(3.18) \quad \begin{aligned} W(\lambda, \alpha, f, x) &= V(\mu) e^{xD(\mu)} U(\mu, \alpha)^{-1} V(\mu) \int_0^{2\pi} e^{(2\pi-s)D(\mu)} V(\mu)^{-1} \varphi(s) ds \\ &\quad + V(\mu) \int_0^x e^{(x-s)D(\mu)} V(\mu)^{-1} \varphi(s) ds \\ &= e^{xF(\lambda)} T(\lambda, \alpha)^{-1} \int_0^{2\pi} e^{(2\pi-s)F(\lambda)} \varphi(s) ds + \int_0^x e^{(x-s)F(\lambda)} \varphi(s) ds, \end{aligned}$$

where (see (3.7) and (3.10))

$$T(\lambda, \alpha) = B(\alpha) - e^{2\pi F(\lambda)}.$$

The boundedness of the last term in (3.18) for λ in a neighborhood of zero is clearly no problem. For the first term after the last equality we note that $e^{xF(\lambda)}$ is bounded, and integrating by parts and using $[f] = 0, f \in L_0^2(0, 2\pi)$, we have

$$T(\lambda, \alpha)^{-1} \int_0^{2\pi} e^{(2\pi-s)F(\lambda)} \varphi(s) ds = T(\lambda, \alpha)^{-1} \hat{F}(\lambda) \int_0^{2\pi} e^{(2\pi-s)F(\lambda)} \Phi(s) ds,$$

wherein $\hat{F}(\lambda) = 2\pi F(\lambda)$ and

$$\Phi(s) = \int_0^s \varphi(\sigma) d\sigma.$$

Clearly then, it is sufficient to show that the matrix $T(\lambda, \alpha)^{-1} \hat{F}(\lambda)$ remains bounded in a neighborhood of $\lambda = 0$. From the easily verified property $F(\lambda)^3 = -\lambda I$, where I is now the 3×3 identity matrix, and from the power series expansion of $e^{F(\lambda)}$ it is clear that

$$(3.19) \quad \hat{F}(\lambda)^3 = -8\pi^3 \lambda I, \quad \hat{F}(\lambda)^{-1} = -\frac{1}{8\pi^3 \lambda} \hat{F}(\lambda)^2.$$

Using the Taylor expansion of $e^{2\pi F(\lambda)} = e^{\hat{F}(\lambda)}$,

$$e^{\hat{F}(\lambda)} = I + \hat{F}(\lambda) + \frac{1}{2} \hat{F}(\lambda)^2 + \frac{1}{6} \hat{F}(\lambda)^3 + \hat{F}(\lambda)^4 G(\lambda),$$

where $G(\lambda)$ is bounded in a neighborhood of $\lambda = 0$, we obtain

$$\begin{aligned} T(\lambda, \alpha)^{-1} \hat{F}(\lambda) &= (B(\alpha) - I - \hat{F}(\lambda) - \frac{1}{2} \hat{F}(\lambda)^2 - \frac{1}{6} \hat{F}(\lambda)^3 - \hat{F}(\lambda)^4 G(\lambda))^{-1} \hat{F}(\lambda) \\ &= (\hat{F}(\lambda)^{-1} (B(\alpha) - I - \hat{F}(\lambda) - \frac{1}{2} \hat{F}(\lambda)^2 - \frac{1}{6} \hat{F}(\lambda)^3 - \hat{F}(\lambda)^4 G(\lambda)))^{-1} \\ &= \left(\frac{1}{8\pi^3 \lambda} \hat{F}(\lambda)^2 (I + \hat{F}(\lambda) + \frac{1}{2} \hat{F}(\lambda)^2 + \frac{1}{6} \hat{F}(\lambda)^3 + \hat{F}(\lambda)^4 G(\lambda) - B(\alpha)) \right)^{-1}, \\ (3.20) \quad &= \left(\frac{1}{8\pi^3 \lambda} \hat{F}(\lambda)^2 (I - B(\alpha) - \frac{4}{3} \pi^3 \lambda I) - I - \frac{1}{2} \hat{F}(\lambda) + 8\pi^3 \lambda G(\lambda) \right)^{-1} \\ &= 2\pi (\lambda^{-1} F(\lambda)^2 (I - B(\alpha) - \frac{4}{3} \pi^3 \lambda I) - 2\pi I - 2\pi^2 F(\lambda) + \mathcal{O}(|\lambda|))^{-1} \\ &= \left(\begin{pmatrix} -1 & -\pi & -\frac{2}{3} \pi^2 \\ \frac{2}{3} \pi^2 \lambda & -1 & -\pi \\ \pi \lambda & \frac{2}{3} \pi^2 \lambda + (\alpha - 1)/2\pi & -1 \end{pmatrix} + \mathcal{O}(|\lambda|) \right)^{-1} \\ &= \left(\begin{pmatrix} -1 & -\pi & -\frac{2}{3} \pi^2 \\ 0 & -1 & -\pi \\ 0 & (\alpha - 1)/2\pi & -1 \end{pmatrix} + \mathcal{O}(|\lambda|) \right)^{-1}, \quad \lambda \rightarrow 0. \end{aligned}$$

Since the indicated constant matrix has determinant $-(1 + \alpha)/2 \neq 0$, we conclude that the matrix (3.20) is uniformly bounded in a neighborhood of $\lambda = 0$ and hence, as developed earlier, that (3.17) is valid and $R(\lambda, \tilde{A}_{\mathcal{H}})$ is, consequently, bounded on the imaginary axis. Applying Huang's cited result and Lemma 4 below, we conclude that the restricted semigroup $\tilde{S}_{\mathcal{H}}(t)$ has the property

$$\|\tilde{S}_{\mathcal{H}}(t)\| \leq M e^{-\gamma t}, \quad t \geq 0,$$

for some positive M and γ , which is equivalent to (3.3). With this the proof of Theorem is complete. \square

LEMMA 4. *We have the estimate*

$$\|R(i\omega, A_{\mathcal{H}})\| = \mathcal{O}(\omega^{-2/3}). \quad |\omega| \rightarrow \infty.$$

Proof. It is clear that w can be expressed in terms of f by using (3.8) and (3.16). It is computationally convenient to recall that this relationship has the form

$$(3.22) \quad w(\lambda, \alpha, x) = \int_0^{2\pi} G(\lambda, x, \xi) f(\xi) d\xi,$$

where for each $\xi \in [0, 2\pi]$ and λ such that (see (3.13)) $\Delta(\mu, \alpha) \neq 0$, $G(\lambda, x, \xi)$ satisfies, for $x \in (0, 2\pi)$,

$$(3.22) \quad G'''(\lambda, x, \xi) + \lambda G(\lambda, x, \xi) = \delta(x - \xi),$$

$$(3.24) \quad G(\lambda, 2\pi, \xi) = G(\lambda, 0, \xi),$$

$$(3.25) \quad G'(\lambda, 2\pi, \xi) = \alpha G'(\lambda, 0, \xi),$$

$$(3.26) \quad G''(\lambda, 2\pi, \xi) = G''(\lambda, 0, \xi),$$

the prime notation now representing d/dx . With μ_0, μ_1, μ_2 as defined earlier, we readily see that $G(\lambda, x, \xi)$ has the form, for coefficients c_0, c_1, c_2 to be determined,

$$(3.27) \quad \begin{aligned} G(\lambda, x, \xi) = & c_0 e^{\mu_0(x-\xi)} + c_1 e^{\mu_1(x-\xi)} + c_2 e^{\mu_2(x-\xi)} \\ & + H(x - \xi)(\hat{c}_0 e^{\mu_0(x-\xi)} + \hat{c}_1 e^{\mu_1(x-\xi)} + \hat{c}_2 e^{\mu_2(x-\xi)}), \end{aligned}$$

where $H(x - \xi)$ is the Heaviside function

$$H(x - \xi) = \begin{cases} 1, & x > \xi, \\ 0, & x \leq \xi, \end{cases}$$

and, corresponding to satisfaction of (3.23),

$$(3.28) \quad \begin{aligned} \hat{c}_0 + \hat{c}_1 + \hat{c}_2 &= 0, \\ \hat{c}_0 \mu_0 + \hat{c}_1 \mu_1 + \hat{c}_2 \mu_2 &= 0, \\ \hat{c}_0 \mu_0^2 + \hat{c}_1 \mu_1^2 + \hat{c}_2 \mu_2^2 &= 1. \end{aligned}$$

One easily computes

$$(3.29) \quad \hat{c}_0 = \frac{1}{(\mu_1 - \mu_0)(\mu_2 - \mu_0)}, \quad \hat{c}_1 = \frac{1}{(\mu_0 - \mu_1)(\mu_2 - \mu_1)}, \quad \hat{c}_2 = \frac{1}{(\mu_0 - \mu_2)(\mu_1 - \mu_2)}.$$

The coefficients c_0, c_1, c_2 are then to be chosen so that (3.24)–(3.26) are satisfied. If the three-dimensional vector whose components are c_k , $k = 0, 1, 2$, and the 3×3 matrix whose diagonal elements are $e^{\mu_k x}$, $k = 0, 1, 2$, are represented by c and $e^{\mu x}$, respectively, these equations take the form (see (3.8) and (3.12))

$$U(\mu, \alpha) e^{-\mu \xi} c = a(\mu, \xi) \equiv V(\mu) e^{\mu(2\pi - \xi)} \hat{c}(\mu),$$

$\hat{c}(\mu)$ being the vector whose components appear in (3.29). Here $U(\mu, \alpha)$ has the form

$$(3.31) \quad U(\mu, \alpha) = \begin{pmatrix} 1 - e^{2\pi \mu_0} & 1 - e^{2\pi \mu_1} & 1 - e^{2\pi \mu_2} \\ \mu_0(\alpha - e^{2\pi \mu_0}) & \mu_1(\alpha - e^{2\pi \mu_1}) & \mu_2(\alpha - e^{2\pi \mu_2}) \\ \mu_0^2(1 - e^{2\pi \mu_0}) & \mu_1^2(1 - e^{2\pi \mu_1}) & \mu_2^2(1 - e^{2\pi \mu_2}) \end{pmatrix}.$$

Solving (3.30), we have

$$(3.32) \quad e^{-\mu \xi} c = \Delta(\mu, \alpha)^{-1} b(\mu, \xi) \equiv \hat{a}(\mu, \xi),$$

where, according to Cramer's rule, $b(\mu, \xi)$ is the three-dimensional vector whose components $b_k(\mu, \xi)$, $k = 0, 1, 2$, are the determinants of the matrices obtained from $U(\mu, \alpha)$ by replacing the k th column of the matrix by $a(\mu, \xi)$. Letting ε be the three-dimensional vector whose components are all equal to 1 and ε^* its (row vector) transpose, we have

$$(3.33) \quad \begin{aligned} G(\lambda, x, \xi) &= \varepsilon^*(e^{\mu(x-\xi)} c + H(x - \xi) e^{\mu(x-\xi)} \hat{c}(\mu)) \\ &= \varepsilon^*(e^{\mu x} \hat{a}(\mu, \xi) + H(x - \xi) e^{\mu(x-\xi)} \hat{c}(\mu)). \end{aligned}$$

Let us take $\omega = \rho^3$, $\rho > 0$. Then the cube roots of $-\lambda = -i\rho^3$ can be taken to be $\mu_0 = i\rho$, $\mu_1 = \rho e^{i7\pi/6}$, $\mu_2 = \rho e^{-i\pi/6}$, which we abbreviate as ρe_0 , ρe_1 , ρe_2 , respectively. As $\rho \rightarrow \infty$, (3.30) yields the asymptotic relationship

$$(3.34) \quad a_k(\mu, \xi) \approx \mu_2^k \hat{c}_2(\mu) e^{\mu_2(2\pi - \xi)}, \quad k = 0, 1, 2.$$

From the form (3.31) of $U(\mu, \alpha)$ it is clear that $\Delta(\mu, \alpha)$, its determinant, has the asymptotic form, again as $\rho \rightarrow \infty$,

$$(3.35) \quad \Delta(\mu, \alpha) \approx -e^{2\pi\mu_2} D(\mu),$$

where

$$\begin{aligned} (3.36) \quad D(\mu) &= \det \begin{pmatrix} 1 - e^{2\pi\mu_0} & 1 & 1 \\ \mu_0(\alpha - e^{2\pi\mu_0}) & \alpha\mu_1 & \mu_2 \\ \mu_0^2(1 - e^{2\pi\mu_0}) & \mu_1^2 & \mu_2^2 \end{pmatrix} \\ &= -\rho^3 \{ (e_0(e_2^2 - e_1^2))(\alpha - e^{2\pi\mu_0}) - \alpha(e_1(e_2^2 - e_0^2))(1 - e^{2\pi\mu_0}) \\ &\quad + (e_2(e_1^2 - e_0^2))(1 - e^{2\pi\mu_0}) \} \\ &= -\rho^3 \{ (e_2 - e_1)((\alpha - 1) + 1 - e^{2\pi\mu_0}) + \alpha(e_2 - e_1)(1 - e^{2\pi\mu_0}) + (e_2 - e_1)(1 - e^{2\pi\mu_0}) \} \\ &= \sqrt{3}\rho^3 \{ (1 - \alpha) - (2 + \alpha)(1 - e^{2\pi\mu_0}) \}. \end{aligned}$$

A simple geometric argument then shows that

$$(3.37) \quad \Delta(\mu, \alpha) \approx \sqrt{3}\rho^3 e^{2\pi\mu_2} \delta(\mu_0, \alpha),$$

where $|\delta(\mu_0, \alpha)| \geq 1 - \alpha > 0$ for all values of ρ in $\mu_0 = \rho i$. In particular, this confirms for large ρ our earlier conclusion that $\Delta(\mu, \alpha)$ does not vanish when λ lies on the imaginary axis.

From the description of $b(\mu, \xi)$ following (3.32) and from (3.34), we further see that, asymptotically as $\rho \rightarrow \infty$,

$$b_k(\mu, \xi) \approx \hat{c}_2(\mu) e^{\mu_2(2\pi - \xi)} H_k(\mu), \quad k = 0, 1, 2,$$

where

$$(3.38) \quad H_0(\mu) = \det \begin{pmatrix} 1 & 1 & 2\pi\mu_2 \\ \mu_2 & \alpha\mu_1 & (\alpha - e^{2\pi\mu_2})\mu_2 \\ \mu_2^2 & \mu_1^2 & (1 - e^{2\pi\mu_2})\mu_2^2 \end{pmatrix} = \det \begin{pmatrix} 1 & 1 & 1 \\ \mu_2 & \alpha\mu_1 & \alpha\mu_2 \\ \mu_2^2 & \mu_1^2 & \mu_2^2 \end{pmatrix},$$

$$(3.39) \quad H_1(\mu) = \det \begin{pmatrix} 1 - e^{2\pi\mu_0} & 1 & 1 - e^{2\pi\mu_2} \\ \mu_0(\alpha - e^{2\pi\mu_0}) & \mu_2 & (\alpha - e^{2\pi\mu_2})\mu_2 \\ \mu_0^2(1 - e^{2\pi\mu_0}) & \mu_2^2 & (1 - e^{2\pi\mu_2})\mu_2^2 \end{pmatrix}$$

$$= \det \begin{pmatrix} 1 - e^{2\pi\mu_0} & 1 & 1 \\ \mu_0(\alpha - e^{2\pi\mu_0}) & \mu_2 & \alpha\mu_2 \\ \mu_0^2(1 - e^{2\pi\mu_0}) & \mu_2^2 & \mu_2^2 \end{pmatrix}.$$

$$(3.40) \quad H_2(\mu) = \det \begin{pmatrix} 1 - e^{2\pi\mu_0} & 1 & 1 \\ \mu_0(\alpha - e^{2\pi\mu_0}) & \alpha\mu_1 & \mu_2 \\ \mu_0^2(1 - e^{2\pi\mu_0}) & \mu_1^2 & \mu_2^2 \end{pmatrix} = D(\mu) \quad (\text{see (3.36)}).$$

Thus we find that as $\rho \rightarrow \infty$

$$\begin{aligned} b_0(\mu, \xi) &\approx \hat{c}_2(\mu) e^{\mu_2(2\pi-\xi)} H_0(\mu), \\ b_1(\mu, \xi) &\approx \hat{c}_2(\mu) e^{\mu_2(2\pi-\xi)} H_1(\mu), \\ b_2(\mu, \xi) &\approx \hat{c}_2(\mu) e^{\mu_2(2\pi-\xi)} D(\mu), \end{aligned} \quad (3.41)$$

so that from (3.32)

$$\hat{a}(\mu, \xi) \approx -\hat{c}_2(\mu) e^{-\mu_2 \xi} \begin{pmatrix} H_0(\mu)/D(\mu) \\ H_1(\mu)/D(\mu) \\ 1 \end{pmatrix} \equiv -\hat{c}_2(\mu) e^{-\mu_2 \xi} h(\mu). \quad (3.42)$$

Therefore, using (3.32) and (3.33), we see that for $f \in L^2(0, 2\pi)$

$$\begin{aligned} G(\lambda, x, \xi) &= \varepsilon^*(e^{\mu(x-\xi)} c + H(x-\xi) e^{\mu(x-\xi)} \hat{c}(\mu)) \\ &= \varepsilon^*(e^{\mu x} \hat{a}(\mu, \xi) + H(x-\xi) e^{\mu(x-\xi)} \hat{c}(\mu)). \end{aligned} \quad (3.43)$$

Since $e^{\mu x} e^{-\mu_2 \xi}$ is uniformly bounded for $x \leq \xi$, the ratios $H_0(\mu)/D(\mu)$ and $H_1(\mu)/D(\mu)$ are bounded as $\rho \rightarrow \infty$ and, as may be clearly seen from (3.29), $|\hat{c}_k(\mu)| \approx \rho^{-2}$ as $\rho \rightarrow \infty$, we conclude that given any $r > 0$ there is a constant \hat{M}_r , independent of ρ , such that for $x \leq \xi$

$$|\varepsilon^* e^{\mu x} \hat{a}(\mu, \xi)| \leq \hat{M}_r \rho^{-2}, \quad \rho > r. \quad (3.44)$$

For $x > \xi$, (3.43) has the asymptotic form

$$G(\lambda, x, \xi) \approx \varepsilon^* e^{\mu(x-\xi)} \hat{c}(\mu) - e^{\mu x} e^{-\mu_2 \xi} \hat{c}_2(\mu) h(\mu).$$

The first two components of this vector involve $e^{\mu_k(x-\xi)}$, $k=0, 1$, and $e^{\mu_k x - \mu_2 \xi}$, $k=0, 1$, all of which are uniformly bounded for $x > \xi$. The third component reduces to zero, as we see from (3.42). We consequently may extend (3.44) to the inequality

$$|G(\lambda, x, \xi)| \leq M_r \rho^{-2}, \quad \rho > r.$$

An entirely similar result is valid for $\lambda = -\rho^3$. Using this and (3.22), we have the result (3.21) and the proof of lemma 4 is complete. \square

COROLLARY 5. *The semigroup $\hat{S}_{\mathcal{H}}(t)$ generated by the operator $A_{\mathcal{H}}$ is a C^∞ semigroup on $L^2(0, 2\pi)$.*

Proof. This result follows immediately from (3.21) and [10, Cor. 4.10]. \square

If we interchange the roles of 2π and 0 in (3.1), equivalently, if we require $\alpha > 1$ in that equation as it stands, the same arguments used above show that the resulting system has uniform exponential decay to the constant state as $t \rightarrow -\infty$. Applying the principle that uniform stabilizability of a time-reversible linear system in both t directions implies exact controllability (see [12], [14], [15]), we conclude that (1.2), (1.3), (1.23) is exactly controllable in $L^2_0(0, 2\pi)$, i.e., controllable between initial and terminal states w_0 and w_T , respectively, for which $[w_T - w_0] = 0$, provided T is large enough so that (see (3.3)) $Me^{-\gamma T} < 1$, using controls $h \in L^2(0, T)$ and keeping the intermediate states $w(\cdot, t)$, $0 \leq t \leq T$, in $L^2(0, 2\pi)$, the norm of h and the maximum norm of $w(\cdot, t)$, $0 \leq t \leq T$, being uniformly bounded in terms of $\|w_0\|_{L^2(0, 2\pi)} + \|w_T\|_{L^2(0, 2\pi)}$; here we effectively reverse the procedure by which we obtained uniform exponential decay in § 2. Harmonic-analysis treatment of the control problem for (1.2), (1.3), (1.23), in much the same manner as for distributed control in § 2, indicates that

we should be able to take $T > 0$ but arbitrarily small. The argument is incomplete, however, because it is apparently not possible to establish that the $L^2(0, T)$ controls $h(t)$ obtained by that route leave the intermediate states $w(\cdot, t)$ in the state space $L^2(0, 2\pi)$ in general—the same difficulty encountered originally in treatment of boundary-value control of the wave equation in more than one space dimension [13], a problem also subsequently overcome by obtaining uniform exponential stabilization rates for the corresponding system with boundary-damping mechanisms [1], [7].

REFERENCES

- [1] G. CHEN, *Energy decay estimates and exact boundary controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–274.
- [2] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators: Part I: General Theory*, Interscience, New York, 1958.
- [3] L. F. HO AND D. L. RUSSELL, *Admissible input elements for systems in Hilbert space and a Carleson-measure criterion*, SIAM J. Control Optim., 21 (1983) pp. 614–640.
- [4] F.-L. HUANG, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–56.
- [5] A. E. INGHAM, *Some trigonometrical inequalities with application to the theory of series*, Math. Z., 41 (1936), pp. 367–379.
- [6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966, p. 495.
- [7] V. KOMORNIK, *Exact controllability in short time for the wave equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 6 (1989), pp. 153–164.
- [8] V. KOMORNIK, D. L. RUSSELL, AND B.-Y. ZHANG, *Control and stabilization of the Korteweg–de Vries equation on a periodic domain*, J. Differential Equations, to appear; Preliminary announcement in C. R. Acad. Sci. Paris, 312 (1991), pp. 841–843.
- [9] G. LUMER AND R. S. PHILLIPS, *Dissipative operators in a Banach space*, Pacific J. Math., 11 (1961), pp. 679–698.
- [10] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [11] D. L. RUSSELL, *Computational study of the Korteweg–de Vries equation with localized control action*, in *Distributed Parameter Control Systems: New Trends and Applications*, G. Chen, E. B. Lee, W. Littman, and L. Markus, eds., Marcel Dekker, New York, 1991.
- [12] —, *Decay rates for weakly damped systems in Hilbert space obtained with control-theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.
- [13] —, *Boundary value control theory of the higher dimensional wave equation, part I*, SIAM J. Control, 9 (1971), pp. 29–42; *part II*, *ibid.*, pp. 401–419.
- [14] —, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in *Differential Games and Control Theory*, E. Roxin, S. Liu, and G. Sternberg, eds., Marcel Dekker, New York, 1974.
- [15] —, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
- [16] B.-Y. ZHANG, *Some results for nonlinear dispersive wave equations with applications to control*, Ph.D. thesis, University of Wisconsin, Madison, June 1990.

DYNAMIC OPTIMIZATION PROBLEMS WITH FREE-TIME AND ACTIVE STATE CONSTRAINTS*

J. D. L. ROWLAND† AND R. B. VINTER†

Abstract. Free-time dynamic optimization problems are treated with state constraints in which the data is permitted to be measurable with respect to the time variable. Necessary conditions for such problems have previously been derived only under the assumption that the state constraint is inactive at the optimal endtimes. A more detailed analysis than has yet been undertaken of the interaction between the optimal free endtimes and the state constraint permits the removal of this assumption. An example clarifies the nature of the new necessary condition.

Key words. necessary conditions, nonsmooth analysis, state constraints, free time

AMS(MOS) subject classifications. 49K15, 49K24

1. Introduction. In recent years, interest has been focused on necessary conditions of optimality for free-time dynamic optimization problems with state constraints, when the data is assumed to be merely measurable (and therefore possibly discontinuous), with respect to the time variable. It has arisen out of the awareness that a variety of threshold phenomena (associated, say, with abrupt changes in a tariff or rate of return on investment at prespecified times) give rise to such problems yet are beyond the scope of traditional necessary condition proof techniques. Our purpose here is to provide a more thorough analysis than has previously been undertaken of the interaction between optimal free endtimes and the state constraint in the “measurable” case. Necessary conditions are thereby provided under significantly weaker hypotheses than have yet been required.

We will soon address a more general problem, but, for the purposes of this introductory discussion, we now trace the development of necessary conditions for the following problem:

$$\begin{aligned}
 &\text{Minimize } g(x(b)) \\
 &\text{over times } b > 0 \text{ and arcs } x(\cdot) \in AC([0, b]; \mathbb{R}^n), \text{ satisfying} \\
 (\mathcal{P}) \quad &\dot{x}(t) \in F(t, x(t)), \quad \text{a.e. } [0, b], \\
 &h(x(t)) \leq 0, \quad \text{for all } t \in [0, b], \\
 &x(0) = x_0.
 \end{aligned}$$

Here $g, h: \mathbb{R}^n \rightarrow \mathbb{R}$ are given functions, $F: \mathbb{R}^{1+n} \rightrightarrows \mathbb{R}^n$ is a given multifunction, and x_0 is a given n -vector.

Necessary conditions for state constrained problems involving a differential inclusion were first obtained for fixed-time problems (problems to which the constraint “ $b = \beta$ ” has been appended, for some fixed $\beta > 0$; see [2]). They lead more or less directly to necessary conditions for free-time problems (\mathcal{P}) via a transformation of the independent variable, which reduces (\mathcal{P}) to a fixed-time problem (see, e.g., [2] or [9]). The following conditions on a minimizer $(\beta, \xi(\cdot))$ are thereby obtained: There

* Received by the editors March 25, 1991; accepted for publication (in revised form) September 16, 1991.

† Department of Electrical Engineering, Imperial College, Exhibition Road, London, SW7 2BT England.

exists an arc $p(\cdot) \in AC([0, b]; \mathbb{R}^n)$, a constant $\lambda \geq 0$, and a nonnegative measure $\mu \in C^*([0, \beta]; \mathbb{R})$, not all zero, satisfying

$$(1.1) \quad (-\dot{p}(t), \dot{\xi}(t)) \in \partial_{x,p} H\left(t, \xi(t), p(t) + \int_{[0,t)} \nabla h(\xi(t)) \mu(dt)\right), \quad \text{a.e. } [0, \beta],$$

$$(1.2) \quad -p(\beta) - \int_{[0,\beta]} \nabla h(\xi(t)) \mu(dt) \in \lambda \nabla g(\xi(\beta)),$$

$$(1.3) \quad \text{supp } \mu \subset \{t: h(\xi(t)) = 0\},$$

$$(1.4) \quad 0 = H\left(\beta, \xi(\beta), p(\beta) + \int_{[0,\beta]} \nabla h(\xi(t)) \mu(dt) - \nabla h(\xi(\beta)) \mu(\{\beta\})\right),$$

where the Hamiltonian $H: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by

$$H(t, x, p) = \sup \{ \langle p, e \rangle : e \in F(t, x) \},$$

and $\partial_{x,p} H$ denotes the (Clarke) generalized gradient of $H(t, \cdot, \cdot)$. For simplicity, we have assumed that the functions $h(\cdot)$ and $g(\cdot)$ are continuously differentiable. Conditions (1.1)–(1.3) will be recognized as the standard fixed-time necessary conditions. Condition (1.4) supplies the extra information associated with the free terminal time.

Unfortunately, this approach is limited to problems where the data is at least Lipschitz continuous with respect to both the time and state variables. The reason is that, under the transformation, the original time variable becomes a component of the state vector, and the fixed-time necessary conditions we would like to apply are valid only for data Lipschitz with respect to the state variable. For problems without state constraints, a more refined analysis generates necessary conditions for free-time problems when the dynamics are continuous in the time variable; see [1], [12], or [13].

When treating problems with data measurable in time, a difficulty arises with the very *interpretation* of the free-time condition (1.4), let alone its proof. In the general setting, it is necessary to specify the data as regards time dependence only in an “almost everywhere” sense, and pointwise evaluation of $t \rightarrow H(t, \xi(t), p(t))$ at $t = \beta$ is meaningless. In [5] this difficulty is overcome, at least for problems without state constraints, by the introduction of the notion of (*convex*) *essential value* of a function.

DEFINITION 1.1. Let $A \subset \mathbb{R}$ be an open set and $h(\cdot): A \rightarrow \mathbb{R}$ be an essentially bounded, measurable function. Take $t \in A$. The (*convex*) *essential value* of $h(\cdot)$ at t , written $\text{co ess}_{s \rightarrow t} h(s)$, is the closed interval with left- and right-hand endpoints

$$\lim_{\epsilon \downarrow 0} \{ \text{ess inf } \{ h(s) : s \in [t - \epsilon, t + \epsilon] \} \}, \quad \lim_{\epsilon \downarrow 0} \{ \text{ess sup } \{ h(s) : s \in [t - \epsilon, t + \epsilon] \} \},$$

respectively. (The definition of the convex essential value coincides with the convex hull of the *essential value* introduced in [5].) It will be clear that adjustments of a function on a null set leave the convex essential value unchanged. Thus the notion of convex essential value is a natural one in the context of necessary conditions. The necessary conditions provided in [5], valid for problems with measurably time-dependent data and when the state constraint is inactive, replace condition (1.4) with the following “essential value” inclusion:

$$0 \in \text{co ess}_{t \rightarrow \beta} H(t, \xi(\beta), p(\beta)).$$

(Note that, when the state constraint is inactive, the measure μ is the zero measure by condition (1.3) and so may be dropped from the conditions.)

Subsequent developments were aimed at providing necessary conditions when the state constraints are active. This was achieved in [4], [15], and [16], but only under the “noninteraction” hypothesis that the state constraints are inactive at the optimal free endtime, or, in other words, only when

$$(1.5) \quad h(\xi(\beta)) < 0.$$

The necessity of investigating the nature of the interaction between the state constraint and the optimal free endtime β is thereby avoided. However, the noninteraction hypothesis (1.5) is unwelcome since it excludes consideration of a frequently encountered phenomenon in these problems, namely, that of a minimizing trajectory evolving on the boundary of the constraint set and terminating at a discontinuity in the dynamics or tariff.

This paper provides an answer to the question of what necessary conditions are valid when the noninteraction hypothesis is dropped. Our necessary conditions retain the flavour of the “Lipschitz case” conditions (1.1)–(1.4) with the exception that the boundary condition on the Hamiltonian (1.4) is replaced by

$$(1.4') \quad 0 = \operatorname{co\,ess}_{t \rightarrow \beta} H\left(t, \xi(\beta), p(\beta) + \int_{[0, \beta]} \nabla h(\xi(t)) \mu(dt) - \kappa \nabla h(\xi(\beta)) \mu(\{\beta\})\right),$$

for some $\kappa \in [0, 1]$. The significant respect in which (1.4') diverges from (1.4) is in the presence of the parameter κ . Its appearance is explained because our proof techniques (in common with those employed in recent research in dynamic optimization) involve extracting “multipliers” $p(\cdot)$, λ , and $\mu(\cdot)$ as limits of multipliers associated with a sequence of perturbed problems. In particular, the measure μ will arise as a limit point in the weak star topology on $\{\mu \in C^*: \mu \geq 0\}$, and so, even if (1.4') is satisfied along the sequence with $\kappa = 1$, we must allow $\kappa \in [0, 1]$ in the limit, due to the fact that the mapping $\mu \rightarrow \mu(\{\beta\})$ is merely weak star upper semicontinuous on the positive cone in C^* .

It is natural to ask whether, when we pass to more general problems than the Lipschitz ones, the need to express the necessary conditions in terms of the parameter κ in the range $0 \leq \kappa \leq 1$ results from a deficient analysis or whether, on the contrary, it represents an essential difference between “Lipschitz” and “non-Lipschitz” problems. The issue is settled by the following example of problem (\mathcal{P}) in which the necessary conditions can be satisfied only if the parameter κ is chosen to lie in $[0, 1]$.

Example 1.2. We consider the following problem involving a two-dimensional state space; state vectors are written (x_1, x_2) , and so forth:

$$\text{Minimize } x_1(b) - x_2(b)$$

over times $b > 0$ and arcs $(x_1(\cdot), x_2(\cdot)) \in AC([0, b]; \mathbb{R}^2)$, satisfying

$$(\dot{x}_1(t), \dot{x}_2(t)) \in \begin{cases} [-1, 1] \times \{1\}, & 0 \leq t \leq 2, \\ \{\eta\} \times \{1\}, & 2 < t, \end{cases} \quad \text{a.e. } t \in [0, b],$$

$$-x_1(t) \leq 0 \quad \text{for all } t \in [0, b],$$

$$(x_1(0), x_2(0)) = (1, 0).$$

Evidently, this is an example of (\mathcal{P}) in which

$$F(t, (x_1, x_2)) = \begin{cases} [-1, 1] \times \{1\}, & 0 \leq t \leq 2, \\ \{\eta\} \times \{1\}, & t > 2 \end{cases},$$

$$h(x_1, x_2) = -x_1,$$

$$g(x_1, x_2) = x_1 - x_2,$$

and the initial state vector is $(1, 0)$. The positive number η is a fixed parameter.

It is obvious that for $\eta < 1$ the example has unbounded cost. On the other hand, if $\eta \geq 1$, the cost is bounded below by -2 . Let us examine the process $(\beta, \xi(\cdot) = (\xi_1(\cdot), \xi_2(\cdot)))$, where $\beta = 2$ and

$$(\xi_1(t), \xi_2(t)) = \begin{cases} (1-t, t), & 0 \leq t \leq 1, \\ (0, t), & 1 \leq t \leq 2. \end{cases}$$

This process has cost -2 . It is a minimizing process if $\eta \geq 1$, but not if $\eta < 1$. We now ask the following question: For what choices of $\lambda \geq 0$, $p(\cdot) = (p_1(\cdot), p_2(\cdot)) \in AC([0, 2]; \mathbb{R}^2)$, and nonnegative measure $\mu \in C^*([0, 2]; \mathbb{R})$, with $\lambda + |p(\cdot)|_\infty + |\mu| > 0$ and $\kappa \in [0, 1]$ are conditions (1.1)–(1.3) and (1.4') satisfied in relation to $(\beta, \xi(\cdot))$?

For this problem, the Hamiltonian function is

$$H(t, (x_1, x_2), (p_1, p_2)) = \begin{cases} |p_1| + p_2, & 0 \leq t \leq 2, \\ p_1 \eta + p_2, & 2 < t. \end{cases}$$

The implications of the Hamiltonian inclusion (1.1) are first explored. Since H is independent of (x_1, x_2) , this condition implies that both $p_1(\cdot)$ and $p_2(\cdot)$ are constant functions. We write

$$p_1(\cdot) = p_1 \quad \text{and} \quad p_2(\cdot) = p_2.$$

Another consequence of (1.1) is that

$$\left\langle p(t) + \int_{[0,t)} \nabla h(\xi(s)) d\mu(s), \dot{\xi}(t) \right\rangle = H\left(t, \xi(t), p(t) + \int_{[0,t)} \nabla h(\xi(s)) d\mu(s)\right), \text{ a.e. } [0, 2].$$

Since $\text{support } \{\mu\} \subset \{h(\xi(t)) = 0\}$, by (1.3), it follows that, for almost every $t \in [0, 1]$,

$$-p_1 + p_2 = |p_1| + p_2.$$

This is only possible if $p_1 \leq 0$. It follows also that, for almost every $t \in [1, 2]$,

$$p_2 = \left| p_1 + \int_{[0,t)} (-1) d\mu(s) \right| + p_2.$$

Since $p_1 \leq 0$ and μ is a nonnegative measure, we conclude that

$$(1.6) \quad p_1 \equiv 0 \quad \text{and} \quad \text{support } \{\mu\} \subset \{2\}.$$

We next study the information embodied in the transversality condition (1.2) and the essential value inclusion (1.4'). In view of (1.6), they assert that

$$(-(-1)\mu(\{2\}), -p_2) = \lambda(1, -1) \quad \text{and} \quad 0 \in \{p_2\} + (1 - \kappa)\mu(\{2\})[-\eta, 1].$$

These combine to tell us that

$$\mu(\{2\}) = p_2 = \lambda \quad \text{and} \quad \lambda \in -\lambda(1 - \kappa)[- \eta, 1].$$

Now we are not permitted to choose $\lambda = 0$, for otherwise these conditions would imply that λ, p_1, p_2 , and μ were all zero, in violation of the requirement that $\lambda + |p(\cdot)|_\infty + |\mu| > 0$. Knowing that $\lambda > 0$, however, we arrive at

$$-(1 - \kappa) \leq 1 \leq (1 - \kappa)\eta.$$

Thus a choice of κ is compatible with (1.1)–(1.3) and (1.4') only if $\kappa \leq 1 - \eta^{-1}$. Evidently, the set of all possible multipliers is

$$\mathcal{M} = \{\lambda > 0, p_1(\cdot) \equiv 0, p_2(\cdot) \equiv \lambda, \mu(\cdot) = \lambda \delta_{\{2\}}(\cdot), \kappa: 0 \leq \kappa \leq 1 - \eta^{-1}\}.$$

Let us now draw conclusions from our characterization of the multiplier set \mathcal{M} . Note first that \mathcal{M} is nonempty if and only if $\eta \geq 1$. Bearing in mind that $(\beta, \xi(\cdot))$ is a minimizing process if and only if $\eta \geq 1$, we see that testing the process $(\beta, \xi(\cdot))$ against the necessary conditions precisely identifies the range of the parameter values η for which $(\beta, \xi(\cdot))$ is a minimizing process. This provides some evidence of the strength of the new necessary conditions. Our second observation is that, in the case where $\eta \geq 1$, when $(\beta, \xi(\cdot))$ is a minimizer, it is not possible to satisfy the necessary conditions with $\kappa = 1$. In fact, if $\eta = 1$, the only choice is $\kappa = 0$! This is consistent with the facts that the differential inclusion is discontinuous in time and that the state constraint is active at the optimal endtime, circumstances when condition (1.4) (the strengthened version of (1.4')) need not apply.

Our discussion has thus far been directed at necessary conditions in the form of a Hamiltonian inclusion for problems having a differential inclusion formulation. However, our methods also provide Pontryagin-type necessary conditions for differential inclusion problems (see § 4) and a general Pontryagin maximum principle for optimal control problems (see § 5). The novelty of these results is, as before, that they apply to problems in which the dynamic constraint is allowed to be merely measurable with respect to the time variable while allowing the state constraint to be active at the optimal endtimes.

We conclude the Introduction by setting our notation for normal cones and generalized differentiation concepts employed in the paper.

DEFINITION 1.3 (normal cones). Let $C \subset \mathbb{R}^n$ be a given closed set and take $x \in C$.

(a) The *proximal normal cone* to C at x , written $\text{PN}_C(x)$, is

$$\text{PN}_C(x) = \{\xi \in \mathbb{R}^n: \text{for some } M > 0, 0 \leq \langle -\xi, c - x \rangle + M|c - x|^2 \quad \forall c \in C\}.$$

(b) The *limiting normal cone*, to C at x , denoted $\text{LN}_C(x)$, is obtained by closing the graph of the multifunction $x \mapsto \text{PN}_C(x)$,

$$\text{LN}_C(x) = \{\xi: \xi_i \rightarrow \xi, x_i \rightarrow x, \xi_i \in \text{PN}_C(x_i), x_i \in C \text{ for all } i\}.$$

Both the proximal and limiting normal cone are subsets of the Clarke normal cone $N_C(x)$, [2]. In fact, the limiting normal cone and Clarke normal cone are related according to

$$N_C(x) = \overline{\text{co}} \text{LN}_C(x).$$

DEFINITION 1.4 (subgradients). Let $\psi: \mathbb{R}^n \rightarrow \mathbb{R}$ be lower semicontinuous near $x \in \mathbb{R}^n$.

(a) The *Clarke generalized gradient* of ψ at x , written $\partial\psi(x)$, is

$$\partial\psi(x) = \{\xi \in \mathbb{R}^n: (\xi, -1) \in N_{\text{epi } \psi}(x, \psi(x))\}.$$

(b) The *limiting subgradient* of ψ at x , written $L\partial\psi(x)$, is

$$L\partial\psi(x) = \{\xi \in \mathbb{R}^n: (\xi, -1) \in \text{LN}_{\text{epi } \psi}(x, \psi(x))\}.$$

For locally Lipschitz functions, the Clarke subgradient and limiting subgradient are related according to the formula $\partial\psi(x) = \text{co } L\partial\psi(x)$.

DEFINITION 1.5 (generalized Jacobians). Let $\psi: \mathbb{R}^n \rightarrow \mathbb{R}^d$ be Lipschitz continuous near $x \in \mathbb{R}^n$. The *generalized Jacobian* of ψ , at x , written $\partial\psi(x)$, is

$$\partial\psi(x) = \left\{ \lim_i \nabla\psi(x_i): x_i \rightarrow x, \psi \text{ is differentiable at } x_i \text{ and } \lim_i \nabla\psi(x_i) \text{ exists} \right\}.$$

There is no ambiguity of notation here since, if $d = 1$ in Definition 1.5 and ψ is Lipschitz continuous near x , the generalized Jacobian and the Clarke generalized gradient coincide.

2. Hamiltonian inclusion necessary conditions. In this section, we supply the main results of the paper. These are necessary conditions of “local” optimality for a generalization (P) of the earlier problem (\mathcal{P}), in which both endtimes are choice variables, namely,

$$(2.1) \quad \begin{aligned} &\text{Minimize } g(a, x(a), b, x(b)) \\ &\text{over intervals } [a, b] \subset \mathbb{R} \text{ and arcs } x(\cdot) \in AC([a, b]; \mathbb{R}^n), \text{ which satisfy} \\ (P) \quad &\dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } [a, b], \\ &h(t, x(t)) \leq 0 \quad \text{for all } t \in [a, b], \\ &(a, x(a), b, x(b)) \in S. \end{aligned}$$

Here $g: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ and $h: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ are given functions, $F: \mathbb{R} \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a given multifunction, and $S \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ is a given set.

An *admissible process* for (P) is a triple $(a, b, x(\cdot))$, comprising left and right endpoints, a and b , respectively, of an interval, and a function $x(\cdot) \in AC([a, b]; \mathbb{R}^n)$, which satisfy the constraints of problem (P). It is convenient to regard the function $x(\cdot)$ in an admissible process $(a, b, x(\cdot))$ as having domain all of \mathbb{R} ; in this event, we extend $x(\cdot)$ to all of \mathbb{R} by constant extrapolation; i.e., we get

$$x(t) = x(a) \quad \text{for all } t < a \quad \text{and} \quad x(t) = x(b) \quad \text{for all } t > b.$$

We also require the concept of an “ ε -neighbourhood” about a given admissible process $(\alpha, \beta, \xi(\cdot))$. Given $\varepsilon > 0$, an admissible process $(a, b, x(\cdot))$ is said to lie in the ε -neighbourhood of $(\alpha, \beta, \xi(\cdot))$, provided that $|x(t) - \xi(t)| < \varepsilon$ for all $t \in \mathbb{R}$, $|a - \alpha| < \varepsilon$, and $|b - \beta| < \varepsilon$. (We employ our extrapolation convention to make sense of this definition.)

An admissible process $(\alpha, \beta, \xi(\cdot))$ is said to be a *local minimizer* for (P) if it achieves the minimum of the functional (2.1) over all admissible processes lying in an ε -neighbourhood of $(\alpha, \beta, \xi(\cdot))$ for some $\varepsilon > 0$. Properties of local minimizers are of interest here, and, consequently, it suffices to invoke hypotheses on the data relating merely to their properties “near” the local minimizer $(\alpha, \beta, \xi(\cdot))$ in question. “Nearness” is quantified by a parameter $\omega \in (0, (\beta - \alpha)/2)$. We employ the notation B for the open unit ball in Euclidean space and denote by $T_\omega(\alpha, \beta, \xi(\cdot))$ the tube

$$T_\omega(\alpha, \beta, \xi(\cdot)) = \{(\tau, \zeta): \tau \in (\alpha - \omega, \beta + \omega), |\zeta - \xi(\tau)| < \omega\}.$$

- (H1) g is Lipschitz continuous on $(\alpha, \xi(\alpha), \beta, \xi(\beta)) + \omega B$.
 (H2) h is continuous on $T_\omega(\alpha, \beta, \xi(\cdot))$; there exists a nonnegative number k_h such that

$$|h(t, x) - h(t, y)| \leq k_h |x - y| \quad \text{for all } (t, x), (t, y) \in T_\omega(\alpha, \beta, \xi(\cdot)).$$

- (H3) $F(\cdot, \cdot)$ is $\mathcal{L} \times \mathcal{B}$ -measurable; $F(t, x)$ is nonempty, compact, and convex-valued on $T_\omega(\alpha, \beta, \xi(\cdot))$; and there exists a nonnegative function $k_F(\cdot) \in L^1$, which is essentially bounded on $(\alpha - \omega, \alpha + \omega) \cup (\beta - \omega, \beta + \omega)$ such that

$$F(t, x) \subset k_F(t)B \quad \text{for all } (t, x) \in T_\omega(\alpha, \beta, \xi(\cdot))$$

and

$$F(t, x) \subset F(t, y) + k_F(t)|x - y|B \quad \text{for all } (t, x), (t, y) \in T_\omega(\alpha, \beta, \xi(\cdot)).$$

(H4) The set S is closed.

The statement of the necessary conditions involves the following notation:

$$\partial_x^+ h(x, t) = \text{co} \left\{ \lim_{i \rightarrow \infty} \gamma_i: \gamma_i \in \partial_x h(t_i, x_i), (t_i, x_i) \rightarrow (t, x), \text{ and } h(t_i, x_i) > 0 \text{ for all } i \right\}.$$

Note that $\partial_x^+ h(t, x) = \emptyset$ at points in $T_\omega(\alpha, \beta, \xi(\cdot))$, where $h(t, x) < 0$, since h is continuous there. If h is smooth and there exist points arbitrarily close to (t, x) at which h is positive, then

$$\partial_x^+ h(t, x) = \nabla_x h(t, x).$$

THEOREM 2.1. *Let $(\alpha, \beta, \xi(\cdot))$ be a local minimizer for (P). Suppose that, for some $\omega > 0$, hypotheses (H1)–(H4) hold. Then there exists an arc $p(\cdot) \in AC([\alpha, \beta]; \mathbb{R}^n)$; numbers k_1 and k_2 ; nonnegative numbers $\lambda \geq 0$, $\rho \geq 0$, and $\sigma \geq 0$; a nonnegative measure $\mu \in C^*([\alpha, \beta]; \mathbb{R})$; and a μ -integrable function $\gamma(\cdot): [\alpha, \beta] \rightarrow \mathbb{R}^n$ such that $\lambda + \|p(\cdot)\|_\infty + \rho + \sigma + \mu([\alpha, \beta]) = 1$ and*

$$(2.2) \quad (-\dot{p}(t), \dot{\xi}(t)) \in \partial_{x,p} H \left(t, \xi(t), p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds) \right) \quad \text{a.e. } [\alpha, \beta],$$

$$(2.3) \quad \begin{aligned} \left(-k_1, p(\alpha), k_2, -p(\beta) - \int_{[\alpha, \beta]} \gamma(t) \mu(dt) \right) &\in \text{LN}_S(\alpha, \xi(\alpha), \beta, \xi(\beta)) \\ &+ \lambda L \partial g(\alpha, \xi(\alpha), \beta, \xi(\beta)) \\ &+ \rho L \partial h(\alpha, \xi(\alpha)) \times \sigma L \partial h(\beta, \xi(\beta)), \end{aligned}$$

$$(2.4) \quad k_1 \in \text{co} \operatorname{ess}_{t \rightarrow \alpha} H(t, \xi(\alpha), p(\alpha)),$$

$$(2.5) \quad k_2 \in \text{co} \operatorname{ess}_{t \rightarrow \beta} H \left(t, \xi(\beta), p(\beta) + \int_{[\alpha, \beta]} \gamma(t) \mu(dt) \right),$$

$$(2.6) \quad \gamma(t) \in \partial_x^+ h(t, \xi(t)) \quad \mu\text{-a.e.},$$

$$(2.7) \quad \text{Supp } \mu \subset \{t \in [\alpha, \beta]: \partial_x^+ h(t, \xi(t)) \neq \emptyset\},$$

$$(2.8) \quad \rho = 0 \quad \text{if } h(\alpha, \xi(\alpha)) < 0, \quad \sigma = 0 \quad \text{if } h(\beta, \xi(\beta)) < 0.$$

The conditions simplify when the constraint “ $h(t, x(t)) \leq 0$ ” is dropped from the formulation of problem (P). This situation can be regarded as an instance of a problem with formulation (P) where $h(\cdot, \cdot) \equiv -1$. (With this choice of function, the state constraint is satisfied by any admissible process.)

THEOREM 2.2. *Let $(\alpha, \beta, \xi(\cdot))$ be a local minimizer for (P). Suppose that, for some $\omega > 0$, hypotheses (H1)–(H4) hold. Suppose further that $h(\cdot, \cdot) \equiv -1$. Then there exists an arc $p(\cdot) \in AC([\alpha, \beta]; \mathbb{R}^n)$, numbers k_1 and k_2 , and a nonnegative number $\lambda \geq 0$ such that $\lambda + \|p(\cdot)\|_\infty = 1$ and*

$$\begin{aligned} &(-\dot{p}(t), \dot{\xi}(t)) \in \partial_{x,p} H(t, \xi(t), p(t)) \quad \text{a.e. } [\alpha, \beta], \\ &(-k_1, p(\alpha), k_2, -p(\beta)) \in \text{LN}_S(\alpha, \xi(\alpha), \beta, \xi(\beta)) + \lambda L \partial g(\alpha, \xi(\alpha), \beta, \xi(\beta)), \\ &k_1 \in \text{co} \operatorname{ess}_{t \rightarrow \alpha} H(t, \xi(\alpha), p(\alpha)), \\ &k_2 \in \text{co} \operatorname{ess}_{t \rightarrow \beta} H(t, \xi(\beta), p(\beta)). \end{aligned}$$

Theorem 2.2 is a straightforward refinement of [3, Cor. 2.2]. It differs only in the respect that here the transversality conditions are expressed in terms of limiting normal cones and limiting subgradients. To accommodate the changes, we must merely note that a transversality condition involving limiting normal cones is actually proved in [4], in the case that $g(a, x, b, y) = \langle \eta, y \rangle$ for some η (although the conclusions are not explicitly stated in this form). The restriction on g is then lifted by means of a state augmentation technique involving the epigraph of g along the lines of that in [4], with the exception that we now exploit the interpretation of subgradients of g in terms of elements in the limiting normal cone to $\text{epi } \{g\}$; see [14].

We stress that a proof of Theorem 2.2 is already available, with the qualifications outlined above, to justify using Theorem 2.2 as a building block in our proof of Theorem 2.1. However, Theorem 2.2 can be regarded as a simple corollary of Theorem 2.1. Indeed, if $h \equiv -1$, then by (2.7) the “state constraint multiplier” μ is the zero measure and so drops out of the conditions; the assertions of Theorem 2.2 follow.

The idea of sharpening the transversality conditions by expressing them in terms of limiting normal cones and subgradients is due to Mordukhovich [11] (although it was not originally proposed in connection with free-time problems with data measurable in t).

Finally, we clarify the relationship between the results of this section and the discussion in the Introduction. Making various simplifying assumptions about the data for problem (P), applying Theorem 2.1, and modifying the state constraint multiplier if necessary by addition of an atom at the right endtime, we obtain the following optimality conditions.

COROLLARY 2.3. *Suppose that the constraint set S is of the form*

$$S = \{0\} \times C \times \mathbb{R} \times \mathbb{R}^n,$$

and that the cost function g and state constraint function h are of the form

$$g(a, x, b, y) = \tilde{g}(y), \quad h(t, x) = \tilde{h}(x),$$

for some set $C \subset \mathbb{R}^n$ and functions $\tilde{g}: \mathbb{R}^n \rightarrow \mathbb{R}$ and $\tilde{h}: \mathbb{R}^n \rightarrow \mathbb{R}$. Let $(\beta, \xi(\cdot))$ be a local minimizer for (P). Suppose that, for some $\omega > 0$, hypotheses (H1)–(H4) hold. In addition, assume that $\tilde{h}(\cdot)$ is continuously differentiable. Then there exist numbers $\lambda \geq 0$, $\kappa \in [0, 1]$, an arc $p(\cdot) \in AC([0, \beta]; \mathbb{R}^n)$, and a nonnegative measure $\mu \in C^([0, \beta]; \mathbb{R}^n)$ such that $\lambda + \|p(\cdot)\|_\infty + \mu([0, \beta]) = 1$,*

$$(-\dot{p}(t), \dot{\xi}(t)) \in \partial_{x,p} H\left(t, \xi(t), p(t) + \int_{[0,t)} \nabla \tilde{h}(\xi(s)) \mu(ds)\right), \quad \text{a.e. } [0, \beta],$$

$$p(0) \in \text{LN}_C(\xi(0)),$$

$$-p(\beta) - \int_{[\alpha, \beta]} \nabla \tilde{h}(\xi(t)) \mu(dt) \in \lambda \text{L}\partial \tilde{g}(\xi(\beta)),$$

$$\text{Supp } \mu \subset \{t \in [0, \beta]: \tilde{h}(\xi(t)) = 0\},$$

$$0 \in \text{co } \text{ess}_{t \rightarrow \beta} H\left(t, \xi(\beta), p(\beta) + \int_{[0, \beta]} \nabla \tilde{h}(\xi(t)) \mu(dt) - \kappa \mu(\{\beta\}) \nabla \tilde{h}(\xi(\beta))\right).$$

3. Proof of Theorem 2.1. Let $(\alpha, \beta, \xi(\cdot))$ and $\omega > 0$ be as stated in the theorem. Define $h^+: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ by $h^+(t, x) = \max\{0, h(t, x)\}$. Since $(\alpha, \beta, \xi(\cdot))$ is a local

minimizer, we may choose some $\delta \in (0, \omega/2)$ such that $(\alpha, \beta, \xi(\cdot))$ is a minimizer for problem $Q(0)$, below

$$\begin{aligned}
 & \text{Minimize } g(a, x(a), b, x(b)) \\
 & \text{over intervals } [a, b] \subset \mathbb{R} \text{ and arcs } x(\cdot) \in AC([a, b]; \mathbb{R}^n), \text{ which satisfy} \\
 & \dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } [a, b], \\
 Q(0) \quad & \int_a^b h^+(t, x(t)) dt \leq 0, \\
 & h(a, x(a)) \leq 0, \quad h(b, x(b)) \leq 0, \\
 & (a, x(a), b, x(b)) \in S, \\
 & |x(\cdot) - \xi(\cdot)|_\infty \leq \delta, \quad |a - \alpha| \leq \delta, \quad |b - \beta| \leq \delta.
 \end{aligned}$$

This is a modification of problem (P), in which candidate arcs are required to lie close to $(\alpha, \beta, \xi(\cdot))$ and in which the original “pointwise” state constraint has been replaced with an equivalent formulation involving a combination of integral and pointwise constraints.

For now, we impose the following additional hypothesis on the data:

(HU) $(\alpha, \beta, \xi(\cdot))$ is the unique minimizer for problem $Q(0)$.

Problem $Q(0)$ is embedded in the family of problems $\{Q(r): r \leq 0\}$, shown below:

$$\begin{aligned}
 & \text{Minimize } g(a, x(a), b, x(b)) \\
 & \text{over intervals } [a, b] \subset \mathbb{R} \text{ and arcs } x(\cdot) \in AC([a, b]; \mathbb{R}^n), \text{ which satisfy} \\
 & \dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } [a, b], \\
 Q(r) \quad & \int_a^b h^+(t, x(t)) dt + r \leq 0, \\
 & h(a, x(a)) \leq 0, \quad h(b, x(b)) \leq 0, \\
 & (a, x(a), b, x(b)) \in S, \\
 & |x(\cdot) - \xi(\cdot)|_\infty \leq \delta, \quad |a - \alpha| \leq \delta, \quad |b - \beta| \leq \delta.
 \end{aligned}$$

Denote by V the value function associated with this family of problems; i.e., $V(r) = \inf Q(r)$ for $r \leq 0$. We adopt the convention of assigning V the value $+\infty$ if the “feasible set” for $Q(r)$ is empty. Sequential compactness arguments such as those employed in the proof of [1, Thm. 3.1.7] permit us to conclude the following lemma.

LEMMA 3.1. Assume hypotheses (H1)–(H4) and (HU).

(a) V is a lower semicontinuous function, and, if $V(r) < +\infty$, then $Q(r)$ has a solution.

(b) Let $\{r_i\}$ be any sequence of numbers increasing to zero such that $V(r_i) < +\infty$ for each i . Let $(a_i, b_i, x_i(\cdot))$ be a minimizer for $Q(r_i)$, $i = 1, 2, \dots$. Then

$$|x_i(\cdot) - \xi(\cdot)|_\infty \rightarrow 0, \quad |a_i - \alpha| \rightarrow 0, \quad |b_i - \beta| \rightarrow 0, \quad \text{as } i \rightarrow \infty.$$

The next step is to give conditions on admissible processes associated with a point in $\text{epi } V$ at which the proximal normal cone is nonempty.

LEMMA 3.2. Let $(v', -\varepsilon')$ be a proximal normal to $\text{epi } V$ at a point (r', e') with $\varepsilon' > 0$. Let $(a', b', x'(\cdot))$ be a minimizer for $Q(r')$ such that

$$|x'(\cdot) - \xi(\cdot)|_\infty \leq \omega', \quad |a' - \alpha| \leq \omega', \quad |b' - \beta| \leq \omega',$$

for some number $\omega' \in (0, \delta)$. Then there exist numbers k'_1 and k'_2 ; nonnegative numbers λ' , ρ' , σ' , and ζ' ; an arc $p'(\cdot) \in AC([a', b']; \mathbb{R}^n)$; and measurable functions $\gamma'(\cdot): [a', b'] \rightarrow \mathbb{R}^n$ and $m'(\cdot): [a', b'] \rightarrow [0, 1]$ such that $\lambda' + \rho' + \sigma' + \|p'(\cdot)\|_\infty + \|m'(\cdot)\|_1 > 0$ and

$$\begin{aligned} (-\dot{p}'(t), \dot{x}'(t)) &\in \partial_{x,p} H\left(t, x'(t), p'(t) + \int_{a'}^t \gamma'(s) \zeta' m'(s) ds\right) \quad \text{a.e. } [a', b'], \\ (-k'_1, p'(a'), k'_2, -p'(b')) &\in \lambda' L \partial g(a', x'(a'), b', x'(b')) + L N_S(a', x'(a'), b', x'(b')) \\ &\quad + \rho' L \partial h(a', x'(a')) \times \sigma' L \partial h(b', x'(b')), \\ k'_1 &\in \text{co ess}_{t \rightarrow a'} H(t, x'(a'), p'(a')), \\ k'_2 &\in \text{co ess}_{t \rightarrow b'} H(t, x'(b'), p'(b')), \\ \gamma'(t) &\in \partial_x^+ h(t, x'(t)) \quad \text{a.e. } \{t \in [a', b']: \partial_x^+ h(t, x'(t)) \neq \emptyset\}, \\ \{t \in [a', b']: m'(t) > 0\} &\subset \{t \in [a', b']: \partial_x^+ h(t, x'(t)) \neq \emptyset\}, \\ \rho' = 0 &\quad \text{if } h(a', x'(a')) < 0, \quad \sigma' = 0 \quad \text{if } h(b', x'(b')) < 0. \end{aligned}$$

Proof. Since the component $-\varepsilon'$ of the proximal normal $(v', -\varepsilon')$ is strictly negative, it must be that the point (r', e') in $\text{epi } V$ takes the form

$$(r', e') = \left(-\int_{a'}^{b'} h^+(t, x'(t)) dt, g(a', x'(a'), b', x'(b')) \right).$$

By definition of the proximal normal $(v', -\varepsilon')$, there exists $M > 0$ such that

$$(3.1) \quad \langle -(v', -\varepsilon'), (r', e') - (r, e) \rangle \leq M |(r, e) - (r', e')|^2$$

for all $(r, e) \in \text{epi } V$.

Take any element $(a, b, x(\cdot))$ that satisfies the constraints of problem $Q(0)$, with the possible exception of the state constraint $\int_{[a,b]} h^+(t, x(t)) dt \leq 0$. Then

$$(r, e) = \left(-\int_a^b h^+(t, x(t)) dt, g(a, x(a), b, x(b)) \right)$$

lies in $\text{epi } V$. Note that, if $r \leq r'$, then the point $(r, e') \in \text{epi } V$. It follows from (3.1) that $\langle -v', r' - r \rangle \leq M |r' - r|^2$ for all $r \leq r'$, whence $v' \geq 0$. In general, we interpret (3.1) as follows: $(a', b', (x'(\cdot), y'(\cdot) \equiv 0))$ is a minimizer for the problem

$$\begin{aligned} &\text{Minimize Max } \{ \mathcal{G}(a, x(a), b, x(b), y(b)), h(a, x(a)), h(b, x(b)) \} \\ &\text{over intervals } [a, b] \subset \mathbb{R} \text{ and arcs } (x(\cdot), y(\cdot)) \in AC([a, b]; \mathbb{R}^{n+1}), \text{ satisfying} \\ &\dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } [a, b], \\ &\dot{y}(t) = h^+(t, x(t)) - h^+(t, x'(t)) \chi_{[a', b']}(t) \quad \text{a.e. } [a, b], \\ &(a, x(a), y(a), b, x(b), y(b)) \\ &\quad \in \{(s_1, s_2, 0, s_4, s_5, s_6): (s_1, s_2, s_4, s_5, s_6) \in S \times \mathbb{R}\}, \\ &|x'(\cdot) - x(\cdot)|_\infty \leq \omega', \quad |a' - a| \leq \omega', \quad |b' - b| \leq \omega'. \end{aligned}$$

Here $\chi_A(t)$ is the indicator function of the set A , and $\mathcal{G}: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$ is the function

$$\begin{aligned} \mathcal{G}(x_1, x_2, x_3, x_4, x_5) &= \varepsilon'(g(x_1, x_2, x_3, x_4) - g(a', x'(a'), b', x'(b'))) + v'(x_5 + m(a, b)) \\ &\quad + |g(x_1, x_2, x_3, x_4) - g(a', x'(a'), b', x'(b'))|^2 + (x_5 + m(a, b))^2, \end{aligned}$$

in which

$$m(a, b) = \int_{a'}^{a' \vee a} h^+(t, x'(t)) dt + \int_b^{b' \vee b} h^+(t, x'(t)) dt.$$

Note that this minimization problem is a dynamic optimization problem without state constraints, and the hypotheses are satisfied under which Theorem 2.2 is applicable. Using the fact that $m(a', b') = 0$ and the rules

$$L\partial(f_1(x) + f_2(x)) \subset L\partial f_1(x) + L\partial f_2(x)$$

and

$$L\partial(f_1(x) \vee f_2(x) \vee f_3(x)) \\ \subset \left\{ \sum_i \alpha_i L\partial f_i(x) : \sum_i \alpha_i = 1, \alpha_i \geq 0, \alpha_i = 0 \text{ if } f_i(x) < (f_1(x) \vee f_2(x) \vee f_3(x)) \right\}$$

governing the calculation of limiting subgradients for locally Lipschitz functions (see, e.g. [11]–[13]), we deduce that there exist nonnegative numbers $\alpha_1, \alpha_2, \alpha_3$, and λ ; numbers h' and k' ; and an arc $q(\cdot) \in AC([a', b']; \mathbb{R}^n)$, such that

$$(3.2) \quad \alpha_1 + \alpha_2 + \alpha_3 = 1, \quad \lambda + |q(\cdot)|_\infty \neq 0,$$

$$\alpha_2 = 0 \quad \text{if } h(a', x'(a')) < 0, \quad \alpha_3 = 0 \quad \text{if } h(b', x'(b')) < 0,$$

$$(3.3) \quad (-\dot{q}(t), \dot{x}'(t)) \in \partial_{x,p} H(t, x'(t), q(t)) - \lambda \alpha_1 v' \partial_x h^+(t, x'(t)) \quad \text{a.e. } [a', b'],$$

$$(3.4) \quad \begin{aligned} (-k'_1, q(a'), k'_2, -q(b')) &\in \lambda \alpha_1 \varepsilon' L\partial g(a', x'(a'), b', x'(b')) + L N_S(a', x'(a'), b', x'(b')) \\ &+ \lambda \alpha_2 L\partial h(a', x'(a')) \times \lambda \alpha_3 L\partial h(b', x'(b')), \end{aligned}$$

$$(3.5) \quad k'_1 \in \text{co ess}_{t \rightarrow a'} H(t, x'(a'), q(a')), \quad k'_2 \in \text{co ess}_{t \rightarrow b'} H(t, x'(b'), q(b')).$$

Define $\lambda' = \lambda \alpha_1 \varepsilon'$, $\rho' = \lambda \alpha_2$, $\sigma' = \lambda \sigma_3$, and $\zeta' = \lambda \alpha_1 v'$. Then λ', ρ', σ' , and $\zeta' \geq 0$, and

$$(3.6) \quad \lambda' + \rho' + \sigma' + |q(\cdot)|_\infty \neq 0,$$

$$(3.7) \quad \rho' = 0 \quad \text{if } h(a', x'(a')) < 0, \quad \sigma' = 0 \quad \text{if } h(b', x'(b')) < 0.$$

Now, arguing as in [4], we deduce from (3.3) existence of measurable functions $\gamma'(\cdot): [a', b'] \rightarrow \mathbb{R}^n$ and $m'(\cdot): [a', b'] \rightarrow [0, 1]$ such that

$$(3.8) \quad (-\dot{q}(t), \dot{x}'(t)) \in \partial_{x,p} H(t, x'(t), q(t)) - \zeta' m'(t) \gamma'(t) \times \{0\} \quad \text{a.e. } [a', b'],$$

$$(3.9) \quad \gamma'(t) \in \partial_x^+ h(t, x'(t)) \quad \text{a.e. } \{t \in [a', b']: \partial_x^+ h(t, x'(t)) \neq \emptyset\},$$

$$(3.10) \quad \{t \in [a', b']: m'(t) > 0\} \subset \{t \in [a', b']: \partial_x^+ h(t, x'(t)) \neq \emptyset\}.$$

Define $p'(\cdot) \in AC([a', b']; \mathbb{R}^n)$ by

$$(3.11) \quad p'(t) = q(t) - \int_{a'}^t \zeta' m'(s) \gamma'(s) ds.$$

It follows from (3.6) and (3.11) that

$$\lambda' + \rho' + \sigma' + |p'(\cdot)|_\infty + |m'(\cdot)|_1 \neq 0.$$

Substituting $(\lambda', \rho', \sigma', p'(\cdot))$ for $(\lambda \alpha_1 \varepsilon', \lambda \alpha_2, \lambda \alpha_3, q(\cdot))$, we arrive at the assertions of the lemma. \square

To conclude the proof of Theorem 2.1 (under the extra hypothesis (HU)), we use the fact that there exist sequences of proximal normals to $\text{epi } V$ and of corresponding base points in $\text{epi } V$, with general terms $(v_i, -\varepsilon_i)$ and (r_i, e_i) , respectively, such that $\varepsilon_i > 0$ along the sequence and $r_i \downarrow 0$. This follows from the lower semicontinuity of V (see [14]). Since $V(r_i) < +\infty$, there exists a minimizer $(a_i, b_i, x_i(\cdot))$ for $Q(r_i)$. By Lemma 3.1,

$$|x_i(\cdot) - \xi(\cdot)|_\infty \rightarrow 0, \quad |a_i - \alpha| \rightarrow 0, \quad |b_i - \beta| \rightarrow 0.$$

By discarding initial terms in the sequences, if necessary, we can arrange that

$$|x_i(\cdot) - \xi(\cdot)|_\infty \leq \bar{\omega}, \quad |a_i - \alpha| \leq \bar{\omega}, \quad |b_i - \beta| \leq \bar{\omega},$$

where $\bar{\omega} \in (0, \delta)$ is a number common to all points along the sequence. Now apply Lemma 3.2. We deduce conditions resembling those in Theorem 2.1, in all respects except that $(a_i, b_i, x_i(\cdot))$ replaces $(\alpha, \beta, \xi(\cdot))$, $(\lambda_i, \rho_i, \sigma_i, p_i(\cdot))$ replaces $(\lambda, \rho, \delta, p(\cdot))$, and $\mu_i(\cdot)$ is defined by

$$\mu_i(dt) = \zeta_i m_i(t) dt$$

replaces μ . We have $\eta_i > 0$, where

$$\eta_i := \lambda_i + \rho_i + \sigma_i + |p_i(\cdot)|_\infty + \mu_i([a_i, b_i]).$$

Scaling the multipliers by η_i^{-1} , we ensure that

$$\lambda_i + \rho_i + \sigma_i + |p_i(\cdot)|_\infty + \mu_i([a_i, b_i]) = 1.$$

Extracting subsequences, we obtain $\lambda, k_1, k_2, \rho, \sigma, p(\cdot)$, and $\mu(\cdot)$ such that $\lambda_i \rightarrow \lambda$, $k_{1i} \rightarrow k_1$, $k_{2i} \rightarrow k_2$, $\rho_i \rightarrow \rho$, $\sigma_i \rightarrow \sigma$, $|p_i(\cdot) - p(\cdot)|_\infty \rightarrow 0$, $\mu_i(\cdot) \xrightarrow{*} \mu(\cdot)$, and

$$\lambda + \rho + \sigma + |p(\cdot)|_\infty + \mu([a, b]) = 1.$$

The assertions of the theorem are proved by passage to the limit, with the help of a convergence analysis along the lines of that in [4] (which uses, in particular, the compactness of solutions to the multifunction F [1, Thm. 3.1.7], upper-semicontinuity properties of the convex essential value [3, Lem. 1.2], and the convergence results of [17]).

It remains to dispose of the extra hypothesis (HU). In circumstances when (HU) is violated, we replace the problem (P) with the problem (\tilde{P}) having cost function

$$g(a, x(a), b, x(b)) + \int_a^b |x(t) - \xi(t)|^2 dt + |b - \beta|^2 + |a - \alpha|^2.$$

The data for problem (\tilde{P}) satisfies the extra hypothesis (H5). Applying the special case of Theorem 2.1 to (\tilde{P}), we deduce the necessary conditions for (P). \square

4. Pontryagin-type conditions. In [8] Kaskosz and Lojasiewicz proved an optimality condition involving a *Carathéodory selection* $f(\cdot, \cdot)$ of the multifunction $F(\cdot, \cdot)$ (defined below) for the fixed-time, state constraint, free version of (P). In their necessary condition, the Hamiltonian inclusion is replaced by a separated adjoint equation and maximum condition reminiscent of the Pontryagin maximum principle. Subsequently, Pontryagin-type conditions were proved by a number of authors in a variety of settings (see [7], [10], [16], and [18]). Pontryagin-type conditions do not subsume Hamiltonian inclusion conditions but supply an independent test of optimality for putative minimizers. (The distinctive features of the two sets of conditions are illustrated by the example in [8].)

Pontryagin-type necessary conditions for problem (P) follow easily from Theorem 2.1, as we demonstrate in this section (cf. [10]).

DEFINITION 4.1. A mapping $f: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is called a Carathéodory selection of the multifunction $F: \mathbb{R} \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ if the single-valued multifunction $\tilde{F}(\cdot, \cdot) := \{f(\cdot, \cdot)\}$ obeys the hypothesis (H3) (with a possibly different Lipschitz constant) and satisfies $f(t, x) \in F(t, x)$ for all $(t, x) \in \mathbb{R} \times \mathbb{R}^n$.

Suppose that the arc $x(\cdot) \in AC([a, b]; \mathbb{R}^n)$ satisfies $\dot{x}(t) \in F(t, x(t))$, almost everywhere $t \in [a, b]$. It is shown in [8] that there always exists a Carathéodory selection f of the multifunction F such that $\dot{x}(t) = f(t, x(t))$, almost everywhere $t \in [a, b]$.

THEOREM 4.2. Let $(\alpha, \beta, \xi(\cdot))$ be a local minimizer for (P). Suppose that, for some $\omega > 0$, hypotheses (H1)–(H4) hold. Pick any Carathéodory selection f of F with the property that

$$\dot{\xi}(t) = f(t, \xi(t)) \quad \text{a.e. } [\alpha, \beta].$$

Then there exists an arc $p(\cdot) \in AC([\alpha, \beta]; \mathbb{R}^n)$; numbers k_1 and k_2 ; nonnegative numbers $\lambda \geq 0$, $\rho \geq 0$, and $\sigma \geq 0$; a nonnegative measure $\mu \in C^*([\alpha, \beta]; \mathbb{R})$; and a μ -integrable function $\gamma(\cdot): [\alpha, \beta] \rightarrow \mathbb{R}^n$ such that $\lambda + \|p(\cdot)\|_\infty + \rho + \sigma + \mu([\alpha, \beta]) = 1$ and

$$\begin{aligned} -\dot{p}(t) &\in \left\langle p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds), \partial_x f(t, \xi(t)) \right\rangle \quad \text{a.e. } [\alpha, \beta], \\ H(t, \xi(t), p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds)) &= \left\langle p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds), \xi(t) \right\rangle \quad \text{a.e. } [\alpha, \beta], \\ \left(-k_1, p(\alpha), k_2, -p(\beta) - \int_{[\alpha, \beta]} \gamma(t) \mu(dt) \right) &\in \text{LN}_S(\alpha, \xi(\alpha), \beta, \xi(\beta)) \\ &\quad + \lambda L \partial g(\alpha, \xi(\alpha), \beta, \xi(\beta)) + \rho L \partial h(\alpha, \xi(\alpha)) \\ &\quad \times \sigma L \partial h(\beta, \xi(\beta)), \\ k_1 &\in \text{co} \operatorname{ess}_{t \rightarrow \alpha} H(t, \xi(\alpha), p(\alpha)), \\ k_2 &\in \text{co} \operatorname{ess}_{t \rightarrow \beta} H(t, \xi(\beta), p(\beta) + \int_{[\alpha, \beta]} \gamma(t) \mu(dt)), \\ \gamma(t) &\in \partial_x^+ h(t, \xi(t)) \quad \mu\text{-a.e.}, \\ \operatorname{Supp} \mu &\subset \{t \in [\alpha, \beta]: \partial_x^+ h(t, \xi(t)) \neq \emptyset\}, \\ \rho = 0 &\quad \text{if } h(\alpha, \xi(\alpha)) < 0, \quad \sigma = 0 \quad \text{if } h(\beta, \xi(\beta)) < 0. \end{aligned}$$

Proof. Take $f(\cdot, \cdot)$ as in the theorem statement. For each $i \in N$, consider the multifunction

$$F_i(t, x) = f(t, x) + i^{-1} \{F(t, x) - f(t, x)\}$$

and the optimization problem (P_i) , obtained from (P) by replacing F with F_i . Note that F_i inherits from F and f the hypotheses (H1)–(H4) for a suitably modified function $k_F(t)$ in (H3). In view of the convexity of $F(t, x)$ and since $f(\cdot, \cdot)$ is a selector of $F(\cdot, \cdot)$, we have $F_i(t, x) \subset F(t, x)$. Now consider the optimization problem (P_i) , obtained from (P) by replacing F with F_i . Since $F_i \subset F$, arcs admissible for (P_i) are also admissible for (P). In view of the fact that $(\alpha, \beta, \xi(\cdot))$ is a local minimizer for (P) and admissible for (P_i) , we deduce that $(\alpha, \beta, \xi(\cdot))$ is also a local minimizer for (P_i) , $i = 1, 2, 3, \dots$. An application of Theorem 2.1 to (P_i) , with reference to $(\alpha, \beta, \xi(\cdot))$,

gives the following information: There exists an arc $p_i(\cdot) \in AC([\alpha, \beta]; \mathbb{R}^n)$; numbers k_1 , and k_{2i} ; nonnegative numbers $\lambda_i \geq 0$, $\rho_i \geq 0$, and $\sigma_i \geq 0$; a nonnegative measure $\mu_i \in C^*([\alpha, \beta]; \mathbb{R})$; and a μ_i -integrable function $\gamma_i(\cdot): [\alpha, \beta] \rightarrow \mathbb{R}^n$ such that $\lambda_i + \|p_i(\cdot)\|_\infty + \rho_i + \sigma_i + \mu_i([\alpha, \beta]) = 1$ and

$$(4.1) \quad (-\dot{p}_i(t), \dot{\xi}(t)) \in \partial_{x,p} H_i\left(t, \xi(t), p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt)\right) \quad \text{a.e. } [\alpha, \beta],$$

$$(4.2) \quad \left(-k_{1i}, p_i(\alpha), k_{2i}, -p_i(\beta) - \int_{[\alpha, \beta]} \gamma_i(t) \mu_i(dt)\right) \in \text{LN}_S(\alpha, \xi(\alpha), \beta, \xi(\beta))$$

$$+ \lambda_i L \partial g(\alpha, \xi(\alpha), \beta, \xi(\beta))$$

$$+ \rho_i L \partial h(\alpha, \xi(\alpha))$$

$$\times \sigma_i L \partial h(\beta, \xi(\beta)),$$

$$(4.3) \quad k_{1i} \in \text{co ess}_{t \rightarrow \alpha} H_i(t, \xi(\alpha), p_i(\alpha)),$$

$$(4.4) \quad k_{2i} \in \text{co ess}_{t \rightarrow \beta} H_i\left(t, \xi(\beta), p_i(\beta) + \int_{[\alpha, \beta]} \gamma_i(t) \mu_i(dt)\right),$$

$$(4.5) \quad \gamma_i(t) \in \partial_x^+ h(t, \xi(t)) \quad \mu_i\text{-a.e.},$$

$$(4.6) \quad \text{Supp } \mu_i \subset \{t \in [\alpha, \beta]: \partial_x^+ h(t, \xi(t)) \neq \emptyset\},$$

$$(4.7) \quad \rho_i = 0 \quad \text{if } h(\alpha, \xi(\alpha)) < 0, \quad \sigma_i = 0 \quad \text{if } h(\beta, \xi(\beta)) < 0,$$

where $H_i(t, x, p) := (1 - i^{-1})\langle p, f(t, x) \rangle + i^{-1}H(t, x, p)$. Expanding (4.1), we obtain

$$(4.8) \quad (-\dot{p}_i(t), \dot{\xi}(t)) \in (1 - i^{-1}) \left[\left\langle p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt), \partial_x f(t, \xi(t)) \right\rangle \times \{f[t, \xi(t)]\} \right]$$

$$+ i^{-1} \partial_{x,p} H\left(t, \xi(t), p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt)\right) \quad \text{a.e. } [\alpha, \beta].$$

Now apply [1, Prop. 3.2.4] to the second component of (4.8) to obtain

$$i^{-1} f(t, \xi(t)) = \dot{\xi}(t) - (1 - i^{-1}) f(t, \xi(t))$$

$$\in i^{-1} \partial_p H\left(t, \xi(t), p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt)\right).$$

It follows that

$$(4.8) \quad H\left(t, \xi(t), p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt)\right)$$

$$= \left\langle p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt), \dot{\xi}(t) \right\rangle \quad \text{a.e. } [\alpha, \beta].$$

Applying the same proposition to the first component of (4.8), we deduce

$$(4.10) \quad -\dot{p}_i(t) \in (1 - i^{-1}) \left\langle p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt), \partial_x f(t, \xi(t)) \right\rangle$$

$$+ i^{-1} \left| p_i(t) + \int_{[\alpha, t)} \gamma_i(t) \mu_i(dt) \right| k_f(t) \bar{B}.$$

Relationships (4.2)–(4.7) and (4.9)–(4.10) are perturbed versions of the assertions of the theorem. Now let $i \rightarrow \infty$. A convergence analysis properties along the lines of that in [4] or [16] completes the proof. \square

5. Optimal control problems. This final section provides necessary conditions for free-time state-constrained dynamic optimization problems having a Pontryagin formulation; i.e., the dynamics are expressed in terms of a differential equation with control term. The novel aspect of these necessary conditions is that, once again, they apply to problems with data merely measurable in the time variable and that they allow the state constraints to be active at the optimal endtimes.

As is customary, the first step is to provide conditions on boundary points of a reachable set. The control system (C) of interest is specified by the following differential equation and constraint:

$$(5.1) \quad (C) \quad \dot{x}(t) = \phi(t, x(t), u(t)), \quad u(t) \in U_t \quad \text{a.e. } [a, b],$$

$$(5.2) \quad h(t, x(t)) \leq 0 \quad \text{for all } t \in [a, b],$$

Here $\phi: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $h: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ are given functions, and $U \subset \mathbb{R} \times \mathbb{R}^m$ is a given set. For $t \in \mathbb{R}$, define $U_t := \{u: (t, u) \in U\}$.

We term *control process* for (C) an element $(a, b, x(\cdot), u(\cdot))$ comprising left and right endpoints, a and b , respectively, of an interval, a function $x(\cdot) \in AC([a, b]; \mathbb{R}^n)$, and a measurable function $u(\cdot): [a, b] \rightarrow \mathbb{R}^m$, which satisfy constraints (5.1) and (5.2). As with inclusion processes, it is convenient to extend the domain of the absolutely continuous function $x(\cdot)$ in a control process to the whole of \mathbb{R} by constant extrapolation. The hypotheses we invoke are of a local nature; they relate to a nominal control process $(\alpha, \beta, \xi(\cdot), v(\cdot))$ and a parameter $\omega > 0$. Let $\varepsilon > 0$ and define the ε -tube of the control process $(\alpha, \beta, \xi(\cdot), v(\cdot))$ by

$$T_\varepsilon(\alpha, \beta, \xi(\cdot), v(\cdot)) = \{(\tau, \eta): \tau \in [\alpha - \varepsilon, \beta + \varepsilon], |\eta - \xi(\tau)| < \varepsilon\}.$$

(HC1) $h(\cdot, \cdot)$ is continuous on $T_\omega(\alpha, \beta, \xi(\cdot), v(\cdot))$; there exists a nonnegative number k_h such that

$$|h(t, x) - h(t, y)| \leq k_h |x - y| \quad \text{for all } (t, x), (t, y) \in T_\omega(\alpha, \beta, \xi(\cdot), v(\cdot)).$$

(HC2) $\phi(\cdot, \cdot)$ is $(\mathcal{L} \times \mathcal{B})$ -measurable for each $x \in \{\eta: (\tau, \eta) \in T_\omega(\alpha, \beta, \xi(\cdot), v(\cdot))\}$. U is $\mathcal{L} \times \mathcal{B}$ -measurable. There exists a nonnegative function $k_\phi(\cdot) \in L^1$, which is essentially bounded on $(\alpha - \omega, \alpha + \omega) \cup (\beta - \omega, \beta + \omega)$ such that

$$|\phi(t, x, u)| \leq k_\phi(t) \quad \text{for all } (t, x) \in T_\omega(\alpha, \beta, \xi(\cdot), v(\cdot)), u \in U_t$$

and

$$|\phi(t, x, u) - \phi(t, y, u)| \leq k_\phi(t) |x - y| B,$$

for all $(t, x), (t, y) \in T_\omega(\alpha, \beta, \xi(\cdot), v(\cdot))$ and $u \in U_t$.

We employ the *unmaximized Hamiltonian* function $\mathcal{H}: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$, defined by

$$\mathcal{H}(t, x, u, p) = \langle p, \phi(t, x, u) \rangle.$$

THEOREM 5.1. *Let $(\alpha, \beta, \xi(\cdot), v(\cdot))$ be a control process for the control system (C) and suppose that hypotheses (HC1) and (HC2) are satisfied with respect to some number $\omega > 0$. Let $C \subset \mathbb{R}^{1+n+1}$ be a closed set and $\psi: \mathbb{R}^n \rightarrow \mathbb{R}^d$ be a function that is Lipschitz*

continuous on a neighbourhood of $(\alpha, \xi(\alpha), \beta)$. Suppose that $(\alpha, \xi(\alpha), \beta) \in C$ and $\psi(\xi(\beta))$ is a boundary point of the reachable set $\mathcal{R}_{\psi, c}^\omega$, defined by

$\mathcal{R}_{\psi, c}^\omega := \{\psi(x(b)): (a, b, x(\cdot), u(\cdot)) \text{ is a control process lying in } T_\omega(\alpha, \beta, \xi(\cdot), v(\cdot))\}$.

Then there exists a function $p(\cdot) \in AC([\alpha, \beta]; \mathbb{R}^n)$; numbers k_1 and k_2 ; nonnegative numbers $\rho \geq 0$ and $\sigma \geq 0$; a vector $\theta \in \mathbb{R}^d$; a nonnegative measure $\mu \in C^*([\alpha, \beta]; \mathbb{R})$; and a μ -integrable function $\gamma(\cdot): [\alpha, \beta] \rightarrow \mathbb{R}^n$ such that $|\theta| + |p(\cdot)|_\infty + \rho + \sigma + \mu([\alpha, \beta]) = 1$ and

$$\begin{aligned} -\dot{p}(t) &\in \partial_x \mathcal{H}\left(t, \xi(t), v(t), p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds)\right) \quad \text{a.e. } [a, b], \\ \mathcal{H}\left(t, \xi(t), v(t), p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds)\right) \\ &= \sup \left\{ \mathcal{H}\left(t, \xi(t), u, p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds)\right) : u \in U_t \right\} \quad \text{a.e. } [a, b], \\ \left(-k_1, p(\alpha), k_2, -p(\beta) - \int_{[\alpha, \beta]} \gamma(t) \mu(dt) \right) &\in \text{LN}_C(\alpha, \xi(\alpha), \beta) \times \partial \psi(\xi(\beta)) \theta \\ &\quad + \rho L \partial h(\alpha, \xi(\alpha)) \times \sigma L \partial h(\beta, \xi(\beta)), \\ k_1 &\in \text{co ess sup}_{t \rightarrow \alpha} \{ \mathcal{H}(t, \xi(\alpha), u, p(\alpha)) : u \in U_t \}, \\ k_2 &\in \text{co ess sup}_{t \rightarrow \beta} \left\{ \mathcal{H}\left(t, \xi(\beta), u, p(\beta) + \int_{[\alpha, \beta]} \gamma(t) \mu(dt)\right) : u \in U_t \right\}, \\ \gamma(t) &\in \partial_x^+ h(t, \xi(t)) \quad \mu\text{-a.e.}, \\ \text{Supp } \mu &\subset \{t \in [\alpha, \beta] : \partial_x^+ h(t, \xi(t)) \neq \emptyset\}, \\ \rho = 0 \quad &\text{if } h(\alpha, \xi(\alpha)) < 0, \quad \sigma = 0 \quad \text{if } h(\beta, \xi(\beta)) < 0. \end{aligned}$$

Consider now the optimal control problem (PC), below:

$$\begin{aligned} &\text{Minimize } \left(g(a, x(a), b, x(b)) + \int_a^b L(t, x(t), u(t)) dt \right) \\ &\text{over intervals } [a, b] \subset \mathbb{R} \text{ and arcs } x(\cdot) \in AC([a, b]; \mathbb{R}^n), \text{ which satisfy} \\ \text{(PC)} \quad &\dot{x}(t) = \phi(t, x(t), u(t)), \quad u(t) \in U_t \quad \text{a.e. } [a, b], \\ &h(t, x(t)) \leq 0 \quad \text{for all } t \in [a, b], \\ &(a, x(a), b, x(b)) \in S, \end{aligned}$$

which is expressed in terms of the control system (C), together with a given set $S \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ and given functions $g: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ and $L: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$. Minimization is conducted over control processes $(a, b, x(\cdot), u(\cdot))$ for system (C) that satisfy the endpoint constraint $(a, x(a), b, x(b)) \in S$. A control process $(a, b, x(\cdot), u(\cdot))$ is said to lie in the ε -neighbourhood of the control process $(a, b, x(\cdot), u(\cdot))$, provided that

$$|a - a| < \varepsilon, \quad |b - b| < \varepsilon, \quad |x(t) - x(t)| < \varepsilon \quad \text{for all } t \in \mathbb{R}.$$

The control process $(a, b, x(\cdot), u(\cdot))$ is a *local minimizer* for (PC) if it achieves the minimum cost over the set of control processes lying in an ε -neighbourhood of

$(a, b, x(\cdot), u(\cdot))$ that satisfy the endpoint constraint. For this “integral cost” problem, the appropriate *unmaximized Hamiltonian* function $\mathcal{H}^\lambda: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$ is

$$\mathcal{H}^\lambda(t, x, u, p) = \langle p, \phi(t, x, u) \rangle - \lambda L(t, x, u).$$

THEOREM 5.2. *Let $(\alpha, \beta, \xi(\cdot), v(\cdot))$ be a local minimizer for (PC). Take $\omega > 0$. Suppose that S is a closed set, $g(\cdot, \cdot, \cdot, \cdot)$ is Lipschitz continuous on a neighbourhood of $(\alpha, \xi(\alpha), \beta, \xi(\beta))$, $h(\cdot, \cdot)$ satisfies hypothesis (HC1), and $\tilde{\phi} (:= \text{column} [\phi, L])$ and U satisfy hypothesis (HC2). Then there exists an arc $p(\cdot) \in AC([\alpha, \beta]; \mathbb{R}^n)$; numbers k_1 and k_2 ; nonnegative numbers $\lambda \geq 0$, $\rho \geq 0$, and $\sigma \geq 0$; a nonnegative measure $\mu \in C^*([\alpha, \beta]; \mathbb{R})$; and a μ -integrable function $\gamma(\cdot); [\alpha, \beta] \rightarrow \mathbb{R}^n$ such that $\lambda + |p(\cdot)|_\infty + \rho + \sigma + \mu([\alpha, \beta]) = 1$ and*

$$\begin{aligned} & -\dot{p}(t) \in \partial_x \mathcal{H}^\lambda \left(t, \xi(t), v(t), p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds) \right) \quad \text{a.e. } [a, b], \\ & \mathcal{H}^\lambda \left(t, \xi(t), v(t), p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds) \right) \\ & = \sup \left\{ \mathcal{H}^\lambda(t, \xi(t), u, p(t) + \int_{[\alpha, t)} \gamma(s) \mu(ds)) : u \in U_t \right\} \quad \text{a.e. } [a, b], \\ & (-k_1, p(\alpha), k_2, -p(\beta) - \int_{[\alpha, \beta]} \gamma(t) \mu(dt)) \in \text{LN}_S(\alpha, \xi(\alpha), \beta, \xi(\beta)) \\ & \quad + \lambda L \partial g(\alpha, \xi(\alpha), \beta, \xi(\beta)) + \rho L \partial h(\alpha, \xi(\alpha)) \\ & \quad \times \sigma L \partial h(\beta, \xi(\beta)), \\ & k_1 \in \text{co ess sup}_{t \rightarrow \alpha} \{ \mathcal{H}^\lambda(t, \xi(\alpha), u, p(\alpha)) : u \in U_t \}, \\ & k_2 \in \text{co ess sup}_{t \rightarrow \beta} \left\{ \mathcal{H}^\lambda(t, \xi(\beta), u, p(\beta) + \int_{[\alpha, \beta]} \gamma(t) \mu(dt)) : u \in U_t \right\}, \\ & \gamma(t) \in \partial_x^+ h(t, \xi(t)) \quad \mu\text{-a.e.}, \\ & \text{Supp } \mu \subset \{t \in [\alpha, \beta] : \partial_x^+ h(t, \xi(t)) \neq \emptyset\}, \\ & \rho = 0 \quad \text{if } h(\alpha, \xi(\alpha)) < 0, \quad \sigma = 0 \quad \text{if } h(\beta, \xi(\beta)) < 0. \end{aligned}$$

Proof of Theorem 5.2. Let $(\alpha, \beta, \xi(\cdot), v(\cdot))$ solve problem (PC). Consider control processes $(a, b, x_1(\cdot), x_2(\cdot), x_3(\cdot), x_4(\cdot), x_5(\cdot), u(\cdot))$, where the state $(x_1(\cdot), x_2(\cdot), x_3(\cdot), x_4(\cdot), x_5(\cdot)) \in AC([a, b]; \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R})$ satisfies dynamical equations

$$(\dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t), \dot{x}_4(t), \dot{x}_5(t)) = (\phi(t, x(t), u(t)), L(t, x(t), u(t)), 0, 0, 0) \quad \text{a.e. } [a, b],$$

$$u(t) \in U_t \quad \text{a.e. } [a, b],$$

$$h(t, x_1(t)) \leq 0 \quad \text{all for } t \in [a, b].$$

Define the Lipschitz continuous function $\psi: \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}$ by

$$\psi(x_1, x_2, x_3, x_4, x_5) = (x_1 - x_3, x_2 - x_4, x_5),$$

the function $G: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$ by

$$G(a, x_1, b, x_3, x_4) = g(a, x_1, b, x_3) + x_4,$$

and the closed set $C \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ by

$$C = \{(a, x_1, x_2, x_3, x_4, b) : (a, x_1, b, x_3) \in S, \quad x_2 = 0, (a, x_1, b, x_3, x_4, x_5) \in \text{epi } G\}.$$

Note that

$$(a, b, x_1(\cdot), x_2(\cdot), x_3(\cdot), x_4(\cdot), x_5(\cdot), u(\cdot)) \\ = \left(\alpha, \beta, \xi(\cdot), \int_{\alpha}^{(\cdot)} L(t, \xi(t), v(t)) dt, \xi(\beta), \int_{\alpha}^{\beta} L(t, \xi(t), v(t)) dt, g(\alpha, \xi(\alpha), \beta, \xi(\beta)) \right. \\ \left. + \int_{\alpha}^{\beta} L(t, \xi(t), v(t)) dt, v(\cdot) \right)$$

is a control process for the five-state control system just defined. Furthermore,

$$(\alpha, x_1(\alpha), x_2(\alpha), x_3(\alpha), x_4(\alpha), x_5(\alpha), \beta) \in C,$$

and, by means of simple contradiction argument, we deduce that, for some $\omega > 0$,

$$\psi(x_1(\beta), x_2(\beta), x_3(\beta), x_4(\beta), x_5(\beta)) \in \text{boundary } \{\mathcal{R}_{\psi, c}^{\omega}\},$$

where $\mathcal{R}_{\psi, c}^{\omega}$ is the reachable set for the five-state control system. We apply Theorem 5.1, and the result follows. \square

Proof of Theorem 5.1. Let $(\alpha, \beta, \xi(\cdot), v(\cdot))$ and ω be as stated in the theorem. The theorem is proved first under the supplementary hypothesis (HCF), below:

(HCF) For each t , the set U_t contains finitely many points.

Take $\delta \in (0, \omega/2)$. Let M be the set of control processes $(a, b, x(\cdot), u(\cdot))$, satisfying

$$|\alpha - a| \leq \delta, \quad |\beta - b| \leq \delta, \quad |\xi(\gamma) - x(\tau)| \leq \delta \quad \text{for } \tau \in [\alpha - \delta, \beta + \delta].$$

Let $a \vee b$ and $a \wedge b$ be the maximum and minimum of a and b , respectively, and consider the metric d on M defined by

$$d((a, b, x(\cdot), u(\cdot)), (a', b', x'(\cdot), u'(\cdot))) = |a - a'| + |b - b'| + |x(0) - x'(0)| \\ + \mathcal{L}\text{-meas } \{t \in [a \vee a', b \wedge b'] : u(t) \neq u'(t)\}.$$

It is a straightforward matter (see [1, Lem. 1, p. 202]) to show that

- (i) M is closed with respect to this metric, and
- (ii) If $(a_i, b_i, x_i(\cdot), u_i(\cdot)) \rightarrow (a, b, x(\cdot), u(\cdot))$ in (M, d) , then

$$\lim_i |x_i(\cdot) - x(\cdot)|_{\infty} = 0, \quad \lim_i a_i = a, \quad \text{and} \quad \lim_i b_i = b.$$

Let $i \in \mathbb{N}$ and let ζ_i be a point in $\psi(\xi(\beta)) + i^{-2}B$ such that $\zeta_i \notin \mathcal{R}_{\psi, c}^{\omega}$. Define the function $\mathcal{G}_i: (M, d) \rightarrow \mathbb{R}$ to be

$$\mathcal{G}_i(a, b, x(\cdot), u(\cdot)) = |\zeta_i - \psi(x(b))|.$$

For each i , \mathcal{G}_i is continuous, and

$$\mathcal{G}_i(\alpha, \beta, \xi(\cdot), v(\cdot)) \leq \inf_{e \in M} \mathcal{G}_i(e) + i^{-2}.$$

By Ekeland's theorem [6], there exists a point $e_i = (a_i, b_i, x_i(\cdot), u_i(\cdot))$ in M such that, writing $\tilde{e} = (\alpha, \beta, \xi(\cdot), v(\cdot))$,

$$d(e_i, \tilde{e}) \leq i^{-1}$$

and

$$\mathcal{G}_i(e_i) \leq \mathcal{G}_i(e) + i^{-1}d(e, e_i) \quad \text{for all } e \in M.$$

By taking i sufficiently large, we can ensure that

$$|\alpha - a_i| < \delta, \quad |\beta - b_i| < \delta, \quad |\xi(\tau) - x_i(\tau)| < \delta \quad \text{for } \tau \in [\alpha - \delta, \beta + \delta].$$

This inequality can then be interpreted as stating that $(a_i, b_i, x_i(\cdot), y_i(\cdot) = \int_{a_i}^{(\cdot)} \Phi_i(t, u_i(t)) dt, u_i(\cdot))$ is a local minimizer for the following optimal control problem:

$$(5.3) \quad \text{Minimize } |\zeta_i - \psi(x(b))| + i^{-1}\{|a - a_i| + |b - b_i| + |x_i(a_i) - x(a)| + y(b)\}$$

over intervals $[a, b] \subset \mathbb{R}$ and arcs $(x(\cdot), y(\cdot)) \in AC([a, b]; \mathbb{R}^{n+1})$, which satisfy

$$(5.4) \quad \dot{x}(t) = \phi(t, x(t), u(t)),$$

$$(5.5) \quad \dot{y}(t) = \Phi_i(t, u(t)), \quad u(t) \in U_i \quad \text{a.e. } [a, b],$$

$$(5.6) \quad h(t, x(t)) \leq 0 \quad \text{for all } t \in [a, b],$$

$$(5.7) \quad (a, x(a), b, y(b)) \in C \times \{0\}.$$

Here $\Phi_i: \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}$ is defined by

$$\Phi_i(t, u) = \begin{cases} 1, & \text{if } t \notin [a_i, b_i], \text{ or } u \neq u_i(t), \\ 0, & \text{otherwise.} \end{cases}$$

Next, we introduce the multifunction $F: \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \rightrightarrows \mathbb{R}^n \times \mathbb{R}$, defined by

$$F(t, x, y) = \{(\phi(t, x, u), \Phi_i(t, u)): u \in U_i\}.$$

Under the hypotheses imposed, the multifunction F obeys hypothesis (H3). (The supplementary hypothesis (HCF) is crucial at this point.) Since the right endpoint is unconstrained, it follows that $(a_i, b_i, x_i(\cdot), y_i(\cdot) = \int_{a_i}^{(\cdot)} \Phi_i(t, u_i(t)) dt, u_i(\cdot))$ remains a local minimizer when the dynamic constraints (5.4) and (5.5) are replaced by the constraint

$$(\dot{x}(t), \dot{y}(t)) \in \text{co } F(t, x(t), y(t)) \quad \text{a.e. } [a, b].$$

We have arrived at a differential inclusion problem of a kind to which the Pontryagin-type necessary conditions are applicable. It follows that, for i sufficiently large, there exists a function $(p_i(\cdot), q_i(\cdot)) \in AC([\alpha - \delta, \beta + \delta]; \mathbb{R}^{n+1})$ (with the family $\{p_i(\cdot)\}$ uniformly bounded and equicontinuous), numbers k_{1i} and k_{2i} , nonnegative numbers $\lambda_i \geq 0$, $\rho_i \geq 0$, and $\sigma_i \geq 0$, a nonnegative measure $\mu_i \in C^*([\alpha - \delta, \beta + \delta]; \mathbb{R})$, a μ_i -integrable function $\gamma_i: [\alpha - \delta, \beta + \delta] \rightarrow \mathbb{R}^n$, and a measurable set $A_i \subset [\alpha - \delta, \beta + \delta]$ such that $\lambda_i + \|(q_i(\cdot), p_i(\cdot))\|_\infty + \rho_i + \sigma_i + \mu_i([\alpha - \delta, \beta + \delta]) = 1$, and we have

$$\mathcal{L}\text{-meas } \{A_i\} \rightarrow \beta - \alpha \quad \text{as } i \rightarrow \infty,$$

for all $t \in A_i$,

$$(5.8) \quad \begin{aligned} -\dot{p}_i(t) &\in \partial_x \mathcal{H}\left(t, x_i(t), u_i(t), p_i(t) + \int_{[a_i, t]} \gamma_i(s) d\mu_i(s)\right), \\ -\dot{q}(t) &= 0, \end{aligned}$$

$$(5.9) \quad \begin{aligned} &\mathcal{H}\left(t, x_i(t), u_i(t), p_i(t) + \int_{[a_i, t]} \gamma_i(s) d\mu_i(s)\right) \\ &\geq \sup_{u \in U_i} \mathcal{H}\left(t, x_i(t), u, p_i(t) + \int_{[a_i, t]} \gamma_i(s) d\mu_i(s)\right) - i^{-1}, \end{aligned}$$

and

$$(5.10) \quad \left(-k_{1i}, p_i(a_i), k_{2i}, -p_i(b_i) - \int_{[a_i, b_i]} \gamma_i(t) d\mu_i(t) \right) \in \text{LN}_C(a_i, x_i(a_i), b_i) \times \partial\psi(x_i(b_i))\theta_i \\ + \rho_i L \partial h(a_i, x_i(a_i)) \\ \times \sigma_i L \partial h(b_i, x_i(b_i)) + i^{-1} B, \\ -q_i(b_i) \in i^{-1} B,$$

$$(5.11) \quad k_{1i} \in \text{co ess} \left\{ \max_{t \rightarrow a_i} \left\{ \max_{u \in U_i} \mathcal{H}(t, x_i, (a_i), u, p_i(a_i)) \right\} \right\} + i^{-1} B,$$

$$(5.12) \quad k_{2i} \in \text{co ess} \left\{ \max_{t \rightarrow b_i} \left\{ \max_{u \in U_i} \left(\mathcal{H}(t, x_i(b_i), u, p_i(b_i)) + \int_{[a_i, b_i]} \gamma_i(s) d\mu_i(s) \right) \right\} \right\} + i^{-1} B,$$

$$(5.13) \quad \gamma_i(t) \in \partial_x^+ h(t, x_i(t)) \quad \mu_i\text{-a.e.},$$

$$(5.14) \quad \text{Supp} \{ \mu_i \} \subset \{ t \in [\alpha - \delta, \beta + \delta] : \partial_x^+ h(t, x_i(t)) \neq \emptyset \},$$

where

$$\theta_i := \lambda_i \frac{\zeta_i - \psi(x_i(b_i))}{|\zeta_i - \psi(x_i(b_i))|},$$

so that

$$(5.15) \quad |\theta_i| + |p_i(\cdot)|_\infty + \rho_i + \sigma_i + \mu_i([\alpha - \delta, \beta + \delta]) \in [1 - i^{-1}, 1 + i^{-1}].$$

Conditions (5.8)–(5.15) will be recognized as perturbed versions of the relations whose validity is asserted in Theorem 5.2. The theorem is proved by subsequence extraction and passage to the limit. Finally, hypothesis (HCF) is removed by application of the technique on p. 207 of [2].

REFERENCES

- [1] L. D. BERKOWITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
- [2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [3] ———, *Methods of Dynamic and Nonsmooth Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.
- [4] F. H. CLARKE, P. D. LOEWEN, AND R. B. VINTER, *Differential inclusions with free time*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1988), pp. 573–593.
- [5] F. H. CLARKE AND R. B. VINTER, *Optimal multiprocesses*, SIAM J. Control Optim., 27 (1989), pp. 1072–1091.
- [6] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.
- [7] H. FRANKOWSKA AND B. KASKOSZ, *A maximum principle for differential inclusions with state constraints*, Systems Control Lett., 11 (1988), pp. 189–194.
- [8] B. KASKOSZ AND S. LOJASIEWICZ, *A maximum principle for generalized control systems*, Nonlinear Anal., 9 (1985), pp. 109–130.
- [9] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
- [10] P. D. LOEWEN AND R. B. VINTER, *Pontryagin-type necessary conditions for differential inclusion problems*, Systems Control Lett., 9 (1987), pp. 263–265.
- [11] B. S. MORDUKHOVICH, *Maximum principle in the problem of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
- [12] ———, *On necessary conditions for an extremum in nonsmooth optimization*, Soviet Math. Dokl., 32 (1985), pp. 215–220.
- [13] ———, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988. (In Russian.)

- [14] R. T. ROCKAFELLAR, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–69.
- [15] J. D. L. ROWLAND AND R. B. VINTER, *A maximum principle for optimal control problems with data discontinuous in time*, IEEE Trans. Automat. Control, 36 (1991), pp. 603–608.
- [16] ———, *Pontryagin type conditions for differential inclusions with free time*, J. Math. Anal. Appl., 165 (1992), pp. 587–597.
- [17] R. B. VINTER AND G. PAPPAS, *A maximum principle for nonsmooth optimal control problems with state constraints*, J. Math. Anal. Appl., 89 (1982), pp. 212–232.
- [18] J. WARGA, *An extension of the Kaskosz maximum principle*, preprint.

OPTIMIZATION OF QUEUES USING AN INFINITESIMAL PERTURBATION ANALYSIS-BASED STOCHASTIC ALGORITHM WITH GENERAL UPDATE TIMES*

EDWIN K. P. CHONG[†] AND PETER J. RAMADGE[‡]

Abstract. Convergence (with probability one) of a stochastic optimization algorithm for a single server queue is proved. The parameter to be optimized is updated using an infinitesimal perturbation analysis estimate of the gradient of the performance measure, and the updates are performed at general times. First, an algorithm in which the parameter is updated before each customer begins service is considered. Then it is shown how this analysis can be extended to an algorithm that updates at certain stopping times. The analysis suggests that the sample path behavior of the algorithm is similar to that of an algorithm that updates at the start of every busy period.

Key words. perturbation analysis, stochastic approximation, queueing

AMS(MOS) subject classifications. 93, 60

1. Introduction. We provide a proof of the convergence of a simple stochastic optimization algorithm based on infinitesimal perturbation analysis (IPA) applied to a single-server queue. Of particular note is that the algorithm can perform a parameter update before each customer starts service. Our convergence proof thus resolves an open problem that has been of interest for some time.

Our results are set in the following framework. Consider a GI/G/1 queue with a performance measure $J(\theta)$, where θ is a control parameter on which the service time distribution depends. Our objective is to find θ^* so that $J(\theta^*)$ is the global minimum value of J . We assume that J has just one local minimum, and this is at θ^* . To achieve our objective we observe a single sample path of the system and use an IPA-based estimate of $dJ/d\theta$ together with a stochastic approximation algorithm to recursively update the value of θ so that $J(\theta)$ is asymptotically minimized. Specifically, we set $\theta_{n+1} = \theta_n - a_n \hat{h}_{n+1}$, where $\{\theta_n\}$ is the sequence of control parameter values, $\{a_n\}$ is a gain sequence, and \hat{h}_{n+1} is an IPA-based estimate of $dJ/d\theta(\theta_n)$. In general, the index of the parameter sequence will be different from the index of the customers in the system, since the parameter updating can be performed less frequently than customers begin service. We assume that the parameter is updated each time a prespecified criterion is met (i.e., according to some stopping time). The key theoretical issue is to prove almost sure convergence of the parameter sequence θ_n to the optimizing value θ^* .

Several authors have examined algorithms within the above framework, e.g., [1]–[8]. In [5] Wardi presented a proof of convergence of an algorithm for the GI/G/1 queue. The task was to minimize $J(\theta) = f(W(\theta))$, where $W(\theta)$ is the mean waiting time, and f is a continuously differentiable function. The algorithm required increasingly larger numbers of customers between parameter updates, and used a nonstandard definition of convergence. In [7] Fu used an ordinary differential equation (ode) analysis to prove the almost sure convergence of an IPA-based algorithm

* Received by the editors March 4, 1991; accepted for publication (in revised form) September 17, 1991.

[†] School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907-1285. This author's research was partially supported by an IBM graduate fellowship.

[‡] Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544. This author's research was partially supported by National Science Foundation grant ECS87-15217.

applied to a GI/G/1 queue with the performance measure $J(\theta) = T(\theta) + C(\theta)$, where $T(\theta)$ is the steady-state mean system time, and $C(\theta)$ is a deterministic cost function on θ . In Fu's algorithm the parameter is updated after every busy period of the queue. In [8] Chong and Ramadge used a martingale-based analysis to prove the almost sure convergence of an IPA-based algorithm applied to an M/G/1 queue with $J(\theta)$ as defined above. This algorithm also updates at the end of each busy period of the queue. These convergence proofs represent a significant step forward. However, the fact that the algorithms require a complete busy period before each update is felt to be unsatisfactory. If the load is high, the busy periods will be very long, and parameter updates will be very infrequent.

Suri and Leung [6] have examined an IPA-based algorithm for an M/M/1 queue in which the parameter updates are performed every time a fixed number of customers have been served. They considered the performance measure $J(\theta) = T(\theta) + c_1/\theta$, where $T(\theta)$ is the mean system time and c_1 is a constant. Their simulation study showed that the algorithm converged in the cases considered, with a relatively fast rate of convergence, corroborating the results in several previous studies (e.g., [2], [4]). Vázquez-Abad [9] has examined a similar scheme for adaptive routing in networks. This form of updating overcomes the restriction of waiting a complete busy period before updating the parameter. Although simulation studies indicate that such algorithms converge, a proof of almost sure convergence has been lacking.

Ideally, we would like to derive expressions for the convergence rates of algorithms involving durations between updates, and thereby be able to compare the convergence rate of an algorithm which updates before the service of each customer with that of one which updates before every busy period. However, this seems to be a difficult task. A necessary first step is to prove the convergence of these algorithms, and this is the focus of our paper. Our main result is a proof of the almost sure convergence of an IPA-based algorithm that updates the parameter at a general sequence of stopping times.¹ However, our approach gives insight not only into why the algorithms converge, but also (intuitively) how fast they should do so in practice. Specifically, our proof indicates a certain similarity of the sample paths of our algorithm with an algorithm that updates at the start of each busy period. This in turn suggests that the rates of convergence of these two algorithms may be very similar.

Our proof relies only on a simple convergence result for stochastic approximation algorithms (Lemma 3.9), rather than building on analogous results in a more sophisticated framework, such as algorithms for Markov chains with transition probabilities that depend on a parameter (see, for example, [11]). As a result our proof introduces some technical arguments that might otherwise have been avoided. We pay this penalty to provide the insight into the sample path behavior of our algorithm as discussed above. The basic method in the proof, however, is fairly simple and easily understood. To aid in the presentation, we focus first on an algorithm that updates the parameter before each customer begins service. Then we extend our analysis to an algorithm that performs parameter updates at certain specified stopping times. For simplicity, we restrict our attention to the performance measure $J(\theta) = T(\theta) + C(\theta)$, where $T(\theta)$ is the mean system time and $C(\theta)$ is a deterministic cost on θ . This choice of J is restrictive, but it simplifies the presentation. We indicate how the analysis can be extended to alternative performance measures in §5.

To prove our main result, it is necessary to impose certain assumptions on our

¹ We are grateful to a reviewer for pointing out that work similar in nature to the problem considered here has recently appeared in [10].

GI/G/1 queue. These are detailed in §3.1. Roughly speaking, the main assumptions are that the service time distribution has finite moments up to a sufficiently high order (at least 6), and that the interarrival time distribution has a bounded hazard rate. We view these assumptions as rather mild. They are satisfied, for example, by exponential, deterministic, and uniform service time distributions, and exponential, Erlang, and hyperexponential interarrival time distributions. Note, however, that our assumption on the interarrival time distribution precludes the uniform distribution.

Although the main contribution of the paper is a proof of convergence of a class of optimization algorithms for certain GI/G/1 queues, we believe our analysis is important in a larger context. To date, all theoretical analyses of IPA-based stochastic optimization of queueing systems have been limited to single-server systems. We believe that the tools and methods of analysis used in our work give additional insight into the behavior of these on-line optimization schemes, and we expect that these methods will be useful in the analysis of optimization algorithms for genuine multiparameter multiqueue systems.

The remainder of this paper is organized as follows. In §2 we describe the probability space over which our system is defined. We also discuss the optimization problem and formulate an IPA-based stochastic optimization algorithm that updates the parameter before the service of each customer. In §3 we prove the almost sure convergence of the algorithm. Extensions of our analysis to more general updating schemes are discussed in §4. In §5 we indicate how our results can be applied to more general performance measures. Section 6 concludes with a discussion of open problems and future work.

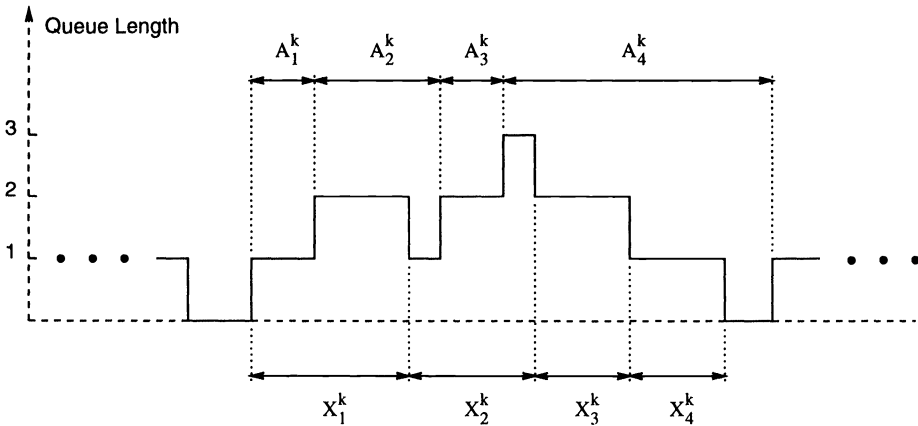
2. Problem formulation.

2.1. Model for a controlled GI/G/1 queue. Consider a GI/G/1 queue with service times that depend on a control parameter θ . We index the customers that arrive at the queue by $n = 1, 2, \dots$. For convenience, we assume that the first customer arrives to find the queue empty. We will use the symbol A_n for the duration of time from the arrival of the n th customer to that of the $(n+1)$ th customer, θ_n for the value of the control parameter for the n th customer, and $X_n(\theta_n)$ for the service time of the n th customer. The queue generates a sequence of alternating *busy periods* (BPs) and *idle periods*. We index the BPs by $k = 1, 2, \dots$, and the customers within each BP by $m = 1, 2, \dots$.

The probability space and corresponding sample path to model the queue can be set up in many ways. Here we will use a particular construction that is convenient for our purposes. We assume that the underlying probability space (Ω, \mathcal{F}, P) is equipped with the following random variables: for each $k \in \mathbb{N}$ (where $\mathbb{N} \triangleq \{1, 2, \dots\}$), sequences $\{A_m^k\}_{m=1}^\infty$ and $\{X_m^k\}_{m=1}^\infty$. The A_m^k are independent over m and k . Each X_m^k is a random positive function $X_m^k : \mathbb{R} \rightarrow \mathbb{R}_+$, and the X_m^k are independent over both m and k , and are also independent of the A_m^k . We write $X_m^k(\theta)$ for the value of the function X_m^k at the argument θ . For every m, k and θ , $X_m^k(\theta)$ is a positive random variable with distribution function $F(x, \theta)$.

We use the random variables defined on (Ω, \mathcal{F}, P) to model our controlled GI/G/1 queue as follows. For each BP k ,

- A_m^k gives the duration between the times of arrival of the m th and $(m+1)$ th customer; and
- $X_m^k(\theta)$ gives the service time of the m th customer in BP k , when the value of the control parameter is θ .

FIG. 1. The k th busy period of the queue.

It remains to specify the control parameter sequence $\{\theta_n\}$. However, once this is done the controlled GI/G/1 queue is completely determined. To see this, let \bar{N}_k denote the number of customers served in the k th BP, and let θ_m^k denote the value of the control parameter for the m th customer in the k th BP. The sequences $\{\bar{N}_k\}$ and $\{\theta_m^k\}$ can be determined recursively from the requirement that \bar{N}_k is the smallest positive integer satisfying

$$\sum_{i=1}^{\bar{N}_k} X_i^k(\theta_i^k) < \sum_{i=1}^{\bar{N}_k} A_i^k.$$

Then the sequence $\{A_n\}$ of interarrival times is given by

$$A_1^1, A_2^1, \dots, A_{\bar{N}_1}^1, A_1^2, A_2^2, \dots$$

and the sequence $\{X_n(\theta_n)\}$ of service times can be written as

$$X_1^1(\theta_1), X_2^1(\theta_2), \dots, X_{\bar{N}_1}^1(\theta_{\bar{N}_1}), X_1^2(\theta_{\bar{N}_1+1}), X_2^2(\theta_{\bar{N}_1+2}), \dots$$

Figure 1 shows a typical BP of the queue.

In contrast to a conventional GI/G/1 queue, the sequence $\{\bar{N}_k\}$ need not be independent nor identically distributed, since the θ values from one BP to the next may be dependent. Indeed, the value of the control parameter θ may differ from customer to customer, so that in contrast to a standard GI/G/1 queue the customer service times will not be identically distributed. In addition, the θ values for different customers will in general be dependent random variables, so the customer service times will also be dependent.

A situation of particular interest is when the control parameter is held constant at a value θ throughout the entire k th BP. In this case the BP is that of a conventional GI/G/1 queue, with the service times independent and identically distributed. Let $N_k(\theta)$ be the number of customers served in the k th BP, assuming that the control parameter is fixed at the value θ throughout this BP. Because of the way we have set up our model we can state that $N_k(\theta)$ is the smallest positive integer satisfying

$$\sum_{i=1}^{N_k(\theta)} X_i^k(\theta) < \sum_{i=1}^{N_k(\theta)} A_i^k.$$

2.2. Stochastic optimization. Our objective is to select the control parameter θ to minimize the performance measure $J(\theta) = T(\theta) + C(\theta)$ with $T(\theta)$ the steady-state mean system time, and $C(\theta)$ a cost function on θ . For example, θ may be the mean service time, and C a decreasing function of θ . To do so we use IPA derivative estimates for $dT(\theta)/d\theta$ together with a stochastic optimization algorithm to recursively update the value of θ .

Since θ will be updated recursively it is important to be precise about what information can be used to perform the updating. For this we introduce the following filtration. Let² $\mathcal{F}_1 = \sigma(\theta_1)$, and for $n > 1$, $\mathcal{F}_n = \sigma(\mathcal{F}_{n-1}, A_{n-1}, X_{n-1}(\theta_{n-1}))$. \mathcal{F}_n represents the information available just before the service of the n th customer. In particular, the service times of all prior customers are known, and the arrival times up to and including the n th customer are known, i.e., these random variables are \mathcal{F}_n -measurable. Note that if $\{f_n\}$ is the random sequence defined by

$$(1) \quad f_n = \begin{cases} 1, & \text{if the } n\text{th customer is the first in a BP,} \\ 0, & \text{otherwise,} \end{cases}$$

then f_n is \mathcal{F}_n -measurable. So, whether a particular customer is the first in a BP is information that can be incorporated into the mechanism for updating θ .

The control parameter sequence is generated by a stochastic optimization algorithm of the form $\theta_{n+1} = \theta_n - a_n \hat{h}_{n+1}$, where θ_n is the control parameter for the service time of customer n , a_n is a step-size, and \hat{h}_{n+1} is an \mathcal{F}_{n+1} -measurable estimate of $dJ/d\theta$. Note in particular that this algorithm updates the control parameter before the service of every customer (general update times are considered in §4).

In general, the control parameter θ_n is subject to some additional constraints. For example, if θ is the mean of the service time, then for stability of the queue, we must have that $0 \leq \theta < E(A_1^1)$. We assume that the parameter values $\{\theta_n\}$ are required to lie in a constraint set D and that D is a compact interval of the real line. To ensure the invariance of D , we modify the update equation to include a *projection*—if $\theta_n \in D$ but $\theta_n - a_n \hat{h}_{n+1} \notin D$, then we set $\theta_{n+1} = \theta_n$ (other forms of projections are also possible). We assume that θ_1 is a random variable taking values in D independent of A_m^k and X_m^k for all m and k (it may indeed be a fixed value). The algorithm is then of the form

$$(2) \quad \theta_{n+1} = \pi_{n+1}[\theta_n - a_n \hat{h}_{n+1}],$$

where $\pi_{n+1}[\cdot]$ is defined by

$$(3) \quad \pi_{n+1}[x] = \begin{cases} \theta_n, & \text{if } x \notin D, \\ x, & \text{otherwise.} \end{cases}$$

It remains to discuss the details of the derivative estimate \hat{h}_{n+1} . This will be done in the next section.

2.3. IPA estimate of $dJ/d\theta$. We begin with some assumptions on $X_m^k(\theta)$ and A_m^k .

Assumptions on $X_m^k(\theta)$ and A_m^k .

(P1) There exists a constant $\theta_{\max} \in D$ such that for all m, k and $\theta \in D$, $X_m^k(\theta) \leq X_m^k(\theta_{\max})$ almost surely;

² We use the notation $\sigma(Y)$ to denote the σ -algebra generated by Y , where Y may be random variable, a σ -algebra, or a combination of both.

(P2) $E(X_1^1(\theta_{\max})) < E(A_1^1)$;

(P3) $X_m^k(\theta)$ is almost surely differentiable with respect to θ on D , and for each $\theta \in D$, there exists a Borel measurable function ψ_θ such that $dX_m^k/d\theta(\theta) = \psi_\theta(X_m^k(\theta))$;

(P4) $E(\sup_{\theta \in D} dX_1^1(\theta)/d\theta)^2 < \infty$;

(P5) $E(X_1^1(\theta_{\max}))^2 < \infty$.

Assumption (P1) ensures that the number of customers served in BP k is bounded above by the random variable $N_k(\theta_{\max})$. For simplicity, we have assumed θ_{\max} to be a constant. This can be relaxed to allow θ_{\max} to be a random variable independent of θ_1 , X_m^k and A_m^k for all m and k . Assumption (P2) ensures that the system is stable for any fixed parameter value $\theta \in D$. Assumptions (P3) and (P4) are standard assumptions (see, for example, [4], [12]), to ensure that the IPA derivative estimates have certain desirable properties. They are generally easy to check. For example, if θ is a so-called scale parameter of $X_m^k(\theta)$, then $X_m^k(\theta) = \theta Y_m^k$ where Y_m^k is a random variable (that does not depend on θ). In this case X_m^k is differentiable, and (P3) holds since $dX_m^k(\theta)/d\theta = Y_m^k = X_m^k(\theta)/\theta$. Moreover, if $E(Y_m^k)^2 < \infty$, then assumption (P4) holds. Assumption (P5) ensures that the number of customers served in a BP has bounded second moments.

The above assumptions allow us to prove the following result.

LEMMA 2.1. *Suppose conditions (P1)–(P5) hold. Then the steady-state mean system time $T(\theta)$ is differentiable on D , and if $T'(\theta) \triangleq dT(\theta)/d\theta$, then for every $k \in \mathbb{N}$,*

$$T'(\theta) = \frac{E(\beta_k(\theta))}{E(N_k(\theta))},$$

where

$$(4) \quad \beta_k(\theta) = \sum_{i=1}^{N_k(\theta)} \sum_{j=1}^i \psi_\theta(X_j^k(\theta)).$$

Proof. This is a special case of a result in [12]. For completeness we provide a proof in Appendix A, based on the method of [12]. \square

Motivated by Lemma 2.1 we formulate an IPA estimate \hat{g}_{n+1} of $T'(\theta_n)$ as follows. Let $\phi_n(\theta_n) = dX_n(\theta_n)/d\theta$. Using (P3), this can be written as $\phi_n(\theta_n) = \psi_{\theta_n}(X_n(\theta_n))$. Define $\{\hat{g}_{n+1}\}$ by

$$(5) \quad \hat{g}_{n+1} = \begin{cases} \phi_n(\theta_n), & \text{if } f_n = 1, \\ \hat{g}_n + \phi_n(\theta_n), & \text{otherwise.} \end{cases}$$

Since $X_n(\theta_n)$ is \mathcal{F}_{n+1} -measurable, and f_n is \mathcal{F}_n -measurable, \hat{g}_{n+1} is \mathcal{F}_{n+1} -measurable.

Let S_k be the random variable defined by $S_1 \triangleq 0$, and for $k > 1$, $S_k \triangleq \sum_{i=1}^{k-1} \bar{N}_i$. S_k is the total number of customers served just prior to the k th BP. So, the m th customer in BP k is the $(S_k + m)$ th customer in the queue. Note that for each k , and each n such that $S_k + 1 \leq n \leq S_k + \bar{N}_k$,

$$(6) \quad \hat{g}_{n+1} = \sum_{i=S_k+1}^n \phi_i(\theta_i) = \sum_{i=S_k+1}^n \psi_{\theta_n}(X_n(\theta_n)),$$

which is similar to the inner sum of (4). Readers familiar with IPA will recognize this expression as the sample derivative of the system time for customer n .

A natural estimate of $dJ(\theta_n)/d\theta$ would be $\hat{g}_{n+1} + dC(\theta_n)/d\theta$. Our actual estimate \hat{h}_{n+1} of $dJ(\theta_n)/d\theta$ is defined recursively, together with a random sequence $\{p_n\}$, as follows: $p_1 = 0$, and for $n > 1$,

$$(7) \quad \hat{h}_{n+1} = \begin{cases} \hat{g}_{n+1} + \frac{dC}{d\theta}(\theta_n), & \text{if } p_n = 0 \text{ or } f_n = 1 \\ \hat{g}_{n+1} + \frac{dC}{d\theta}(\theta_n) + \hat{h}_n, & \text{otherwise} \end{cases}$$

$$(8) \quad p_{n+1} = \begin{cases} 1, & \text{if } \theta_n - a_n \hat{h}_{n+1} \notin D \\ 0, & \text{otherwise.} \end{cases}$$

The sequence $\{p_n\}$ simply indicates if a projection was used in the determination of θ_n . If no projection was involved in determining θ_n , i.e., $p_n = 0$, then we set $\hat{h}_{n+1} = \hat{g}_{n+1} + dC(\theta_n)/d\theta$, as expected. The same is true, regardless of the value of p_n , if the n th customer (the previous customer) was the first in a BP. However, in the case of all other projections, we adjust the value of \hat{h}_{n+1} so that it also incorporates the value of \hat{h}_n . Effectively, a sequence of projections results in an accumulation of the corresponding \hat{h}_n , subject to the above proviso that this accumulation process is “reset” after the first customer in a BP.

The reason for this seemingly complicated formulation of the estimate will become clear as the development proceeds. Suffice it to say at this point that as indicated above the “nominal” estimate is indeed $\hat{g}_{n+1} + dC(\theta_n)/d\theta$. The accumulation of the estimate over sequences of projections is technically advantageous in our proof and is used to extend the algorithm to deal with more general update times in §4. The resetting of the accumulations after the first customer in each BP is related to our basic method of analysis. Neither of these complications may be necessary in practice. For example, in the algorithm used by Suri and Leung [6], the projection method does not use the accumulation mechanism described above. This is essentially the only difference between our algorithm and the one in [6].

3. Convergence of the algorithm. To analyze the convergence of the algorithm described in the previous section it is necessary to introduce additional assumptions. We begin with a list and brief discussion of these assumptions.

3.1. Assumptions.

Assumptions on $\{a_n\}$.

(G1) $\{a_n\}$ is a sequence of positive random variables adapted to the filtration $\{\mathcal{F}_n\}$;

(G2) $\{a_n\}$ is nonincreasing, i.e., for all n , $a_{n+1} \leq a_n$ almost surely;

(G3) There exist a constant δ , $\frac{1}{2} < \delta \leq 1$, and constants $\bar{A} < \infty$, $\underline{A} > 0$ such that for each n , $\underline{A}n^{-\delta} \leq a_n \leq \bar{A}n^{-\delta}$ almost surely;

(G4) There exists a finite constant B_a such that for all n , $(1/a_{n+1}) - (1/a_n) \leq B_a$ almost surely.

The above assumptions are fairly standard and are generally easy to verify. Note that we require a_n to be \mathcal{F}_n -measurable.

We next place additional assumptions on the $X_m^k(\theta)$ and A_m^k . For this, let $0 < d < 1$ be such that

$$(9) \quad \sum_{n=1}^{\infty} a_n^{1+d} < \infty \quad \text{a.s.}$$

Such a d always exists; any d with $1/\delta - 1 < d < 1$ will do, e.g., $d = 1/(2\delta)$. Then let

$$p = \min\{n \in \mathbb{N} : n \geq 2(1+d)(3+d)/(1-d)\}.$$

Note that $p \geq 6$. In the following, it does not matter which m and k we are referring to, since for fixed θ , the $X_m^k(\theta)$ are identically distributed, and likewise the A_m^k .

Assumptions on $X_m^k(\theta)$.

(S1) $E(X_m^k(\theta_{\max}))^p < \infty$;

(S2) Let $\phi_m^k(\theta)$ denote the random variable $dX_m^k(\theta)/d\theta (= \psi_\theta(X_m^k(\theta)))$, and let $\bar{\phi}_m^k = \sup_{\theta \in D} \phi_m^k(\theta)$. Then, $E(\bar{\phi}_1^1)^p < \infty$;

(S3) For all m, k , there exists a measurable function $F_K : \mathbb{R}^D \rightarrow \mathbb{R}$ such that if $K_m^k = F_K(X_m^k)$, then for all $\theta_1, \theta_2 \in D$, $|\phi_m^k(\theta_2) - \phi_m^k(\theta_1)| \leq K_m^k |\theta_2 - \theta_1|$ and $E(K_1^1)^2 < \infty$;

(S4) The distribution of A_m^k has a bounded hazard rate $\mu(t)$, i.e., there exists a constant $\mu > 0$ such that $\mu(t) \leq \mu$ for all $t \geq 0$.

Assumption (S1), together with assumption (P2), is used to ensure that the p th moment of the number served in a BP for a fixed $\theta \in D$ is bounded (see Lemma 3.3, §3.3). Note that (S1) implies (P5). Assumption (S2) is used to ensure boundedness of the moments of the sample gradient ϕ_m^k , and it also ensures that moments of the random variables $\sum_{i=1}^{N_k(\theta)} \bar{\phi}_i^k$ are bounded (see Lemma 3.4, §3.3). Note that assumption (S2) implies (P4). Assumption (S3) is a Lipschitz condition on the derivative $dX_m^k/d\theta (= \phi_m^k)$, and we assume that the Lipschitz constant K_m^k has finite second moment.

The *hazard rate* of the distribution $F_A(x)$ of A_m^k is a positive function $\mu(t)$ satisfying $F_A(x) = 1 - \exp(-\int_0^x \mu(t)dt)$. If $F_A(x)$ has a density, $f_A(x)$, then it has a hazard rate given by $\mu(t) = f_A(t)/(1 - F_A(t))$. Assumption (S4) is used to ensure that the idle periods are not too short, in the sense that if \bar{I} is the duration of the idle period, then $E(\bar{I}^{-q}) < \infty$ for all $q < 1$ (see Lemma 3.6). This assumption holds for many distributions of interest, including the exponential, Erlang and hyperexponential distributions.

Assumptions (S1) and (S2) are generally easy to check. For example, suppose, as before, that θ is a scale parameter of $X_m^k(\theta)$, i.e., $X_m^k = \theta Y_m^k$. Let $D = [\theta_a, \theta_b]$, and $\theta_{\max} = \theta_b$. If $E(Y_m^k)^p < \infty$, then assumptions (S1) and (S2) hold. In fact, in this case (S3) can also be seen to hold, since $\phi_m^k = Y_m^k$ is independent of θ . In general, however, (S3) may not be easy to check directly. The following proposition gives easily checked conditions which imply (S3).

PROPOSITION 3.1. *Suppose (S2) holds. Suppose that for all $x \geq 0$ and all $\theta_1, \theta_2 \in D$, $|\psi_{\theta_2}(x) - \psi_{\theta_1}(x)| \leq K_1(x)|\theta_2 - \theta_1|$ and for all $\theta \in D$ and all $x_1, x_2 \geq 0$, $|\psi_\theta(x_2) - \psi_\theta(x_1)| \leq K_2(\theta)|x_2 - x_1|$. Let $\bar{K}_m^k \triangleq \sup_{\theta \in D} K_1(X_m^k(\theta))$ and $\bar{L} \triangleq \sup_{\theta \in D} K_2(\theta)$. If $E(\bar{K}_m^k)^2 < \infty$ and $\bar{L} < \infty$, then condition (S3) holds.*

Proof. For the proof, see Appendix B. \square

Finally, we make certain assumptions on the cost function C , which appears in the performance measure J .

Assumptions on $C(\theta)$.

(C1) $C(\theta)$ is differentiable with respect to θ on D ;

(C2) Let $C'(\theta) \triangleq dC(\theta)/d\theta$. Then $C'(\cdot)$ is Lipschitz continuous on D , i.e., there exists a constant $C_M'' < \infty$ such that for all $\theta_1, \theta_2 \in D$, $|C'(\theta_2) - C'(\theta_1)| \leq C_M'' |\theta_2 - \theta_1|$;

(C3) $dJ(\theta_n)/d\theta = 0$ only if $\theta = \theta^*$ for some unique $\theta^* \in \overset{\circ}{D}$ (where $\overset{\circ}{D}$ denotes the interior of D), and θ^* minimizes J on D .

Assumptions (C1)–(C3) are natural assumptions to impose on $C(\theta)$. Assumptions (C1) and (C2) are used to ensure that J is sufficiently smooth, while (C3) ensures the uniqueness of the point θ^* that minimizes J . In fact, (C1) and (C2) together with Lemma 2.1 imply that the objective function J is continuously differentiable on D . Assumption (C3) is strong and can be weakened; however, we adopt it here for simplicity of presentation.

3.2. Main result. We are now ready to state our main result.

THEOREM 3.2. *Suppose that (P1)–(P3), (G1)–(G4), (S1)–(S4), and (C1)–(C3) are satisfied. Then, for the sequence $\{\theta_n\}$ defined by (2), (5), (7), and (8), we have that $\theta_n \rightarrow \theta^*$ as $n \rightarrow \infty$ with probability one.*

The proof of the theorem involves several steps and requires a variety of upper bounds. We present these bounds as a sequence of auxiliary lemmas, then proceed to the main steps in the proof.

3.3. Auxiliary lemmas. Throughout the remainder of the paper, we will use the notation $E_X(Y)$ to denote the conditional expectation of the random variable Y given X (where X may be a σ -algebra or a random variable).

We use assumptions (P2) and (S1) to show that the moments up to order p of N_k are bounded.

LEMMA 3.3. *Suppose that (P2) and (S1) hold. Then there exists a finite constant B_N such that for all $k \in \mathbb{N}$, and $r \leq p$, $E(N_k(\theta_{\max}))^r \leq B_N$.*

Proof. This is a special case of [13, Thm. 1(a)]. The result can also be proved using [14, Thm. 3.1(i); p. 78]. \square

Next we use assumption (S2) to show boundedness of the moments of the random variables $\sum_{i=1}^{N_k(\theta)} \bar{\phi}_i^k$.

LEMMA 3.4. *Suppose (S2) holds. Then there exists a finite constant B_ϕ such that for all $k \in \mathbb{N}$, $\theta \in D$, and $r \leq p$,*

$$E \left(\sum_{i=1}^{N_k(\theta)} \bar{\phi}_i^k \right)^r \leq B_\phi.$$

Proof. For the proof, see Appendix C. \square

Note that since C' is continuous (by (C2)), and D is compact, then C' is bounded on D . Let $C'_M \triangleq \sup_{\theta \in D} |C'(\theta)| < \infty$. Then by Lemma 3.4 there exists a finite constant B'_ϕ such that for all k , $\theta \in D$, and $r \leq p$,

$$(10) \quad E \left(\sum_{i=1}^{N_k(\theta)} \bar{\phi}_i^k + C'_M \right)^r \leq B'_\phi.$$

We next use the assumption on the finiteness of the second moment of the Lipschitz constant K_m^k to show that the sum $\sum_{i=1}^{N_k(\theta)} K_i^k$ has a finite second moment.

LEMMA 3.5. *Suppose (S3) holds. Then there exists a finite constant B_K such that for all $k \in \mathbb{N}$ and $\theta \in D$,*

$$E \left(\sum_{i=1}^{N_k(\theta)} K_i^k \right)^2 \leq B_K.$$

Proof. For the proof, see Appendix D. \square

Finally we use assumption (S4) to show that the idle periods have the required property described in §3.1.

LEMMA 3.6. *Suppose (S4) holds. Let $\{\theta_n\}$ be a sequence of random variables adapted to $\{\mathcal{F}_n\}$, taking values in D . Let*

$$\bar{I} = \sum_{i=1}^{\bar{N}_1} A_i^1 - \sum_{i=1}^{\bar{N}_1} X_i^1(\theta_i^1).$$

Then for each $0 \leq q < 1$ there exists a constant B_I such that $E_{\theta_1}(\bar{I}^{-q}) \leq B_I$.

Proof. We may write $\bar{I} = A_{\bar{N}_1}^1 - L$, where $L = \sum_{i=1}^{\bar{N}_1} X_i^1(\theta_i^1) - \sum_{i=1}^{\bar{N}_1-1} A_i^1$. Note that $0 < L < A_{\bar{N}_1}^1$ almost surely. Now, the conditional distribution of \bar{I} given L is $F_{I|L}(x) = 1 - \exp(-\int_0^x \mu(t+L)dt)$, with a density $f_{I|L}(x) = \mu(x+L) \exp(-\int_0^x \mu(t+L)dt)$. By (S4), there exists a constant $\mu > 0$ such that for all $t \geq 0$, $\mu(t) \leq \mu$. Therefore, $f_{I|L}(x) \leq \mu$. Hence,

$$\begin{aligned} E_{\theta_1}(\bar{I})^{-q} &= E_{\theta_1} \left(\int_0^\infty x^{-q} f_{I|L}(x) dx \right) \\ &= E_{\theta_1} \left(\int_0^1 x^{-q} f_{I|L}(x) dx + \int_1^\infty x^{-q} f_{I|L}(x) dx \right) \\ &\leq E_{\theta_1} \left(\int_0^1 x^{-q} \mu dx + \int_1^\infty f_{I|L}(x) dx \right) \\ &\leq \frac{\mu}{1-q} + 1, \end{aligned}$$

which completes the proof. \square

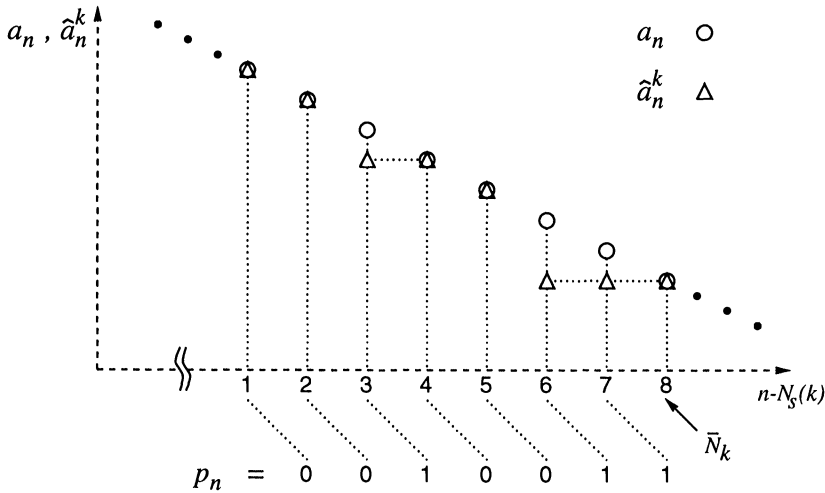
3.4. A technical lemma. In the proof of Theorem 3.2, it will be convenient to relate the terms of the sequence $\{\theta_n\}$ for different customers within the same BP. To this end we define, for each BP k , the random sequence $\{\hat{a}_n^k : S_k + 1 \leq n < \infty\}$ *sample-pathwise* as follows: For $n \geq S_k(\omega) + \bar{N}_k(\omega)$, $\hat{a}_n^k(\omega) = a_{S_k + \bar{N}_k}(\omega)$, and for $S_k(\omega) + 1 \leq n < S_k(\omega) + \bar{N}_k(\omega)$,

$$(11) \quad \hat{a}_n^k(\omega) = \begin{cases} a_n(\omega), & \text{if } p_{n+1}(\omega) = 0, \\ \hat{a}_{n+1}^k(\omega), & \text{otherwise.} \end{cases}$$

Recall that S_k is the total number of customers served just prior to the k th BP. For $n \geq S_k(\omega) + \bar{N}_k(\omega)$, the value of $\hat{a}_n^k(\omega)$ is $\hat{a}_{S_k + \bar{N}_k}(\omega)$, i.e., the value of $a_n(\omega)$ used to update the first customer in the next BP. For $S_k(\omega) + 1 \leq n < S_k(\omega) + \bar{N}_k(\omega)$, the $\hat{a}_n^k(\omega)$ are defined recursively by moving backward through the BP. For $n = S_k(\omega) + \bar{N}_k(\omega) - 1$, the value of $\hat{a}_n^k(\omega)$ is $a_n(\omega)$ if $\theta_{n+1}(\omega)$ is computed without a projection, i.e., $p_{n+1}(\omega) = 0$; otherwise $\hat{a}_n^k(\omega)$ assumes the value of $\hat{a}_{n+1}^k(\omega)$. The value of $\hat{a}_n^k(\omega)$, for $n = S_k(\omega) + \bar{N}_k(\omega) - 2$, is then determined from the values of $a_n(\omega)$, $\hat{a}_{n+1}^k(\omega)$, and $p_{n+1}(\omega)$, and so on. Figure 2 gives an illustration of how the sample paths of $\{\hat{a}_n^k\}$ and $\{a_n\}$ are related. It is clear that \hat{a}_n^k is not \mathcal{F}_n -measurable. However, this will not be of importance in what follows.

By the construction of $\{\hat{a}_n^k\}$ it is easy to see that $\{\hat{a}_n^k\}$ is nonincreasing and that for every k and almost all $\omega \in \Omega$:

- (1) $\hat{a}_n^k(\omega) \geq a_{S_k + \bar{N}_k}(\omega)$;
- (2) For $S_k(\omega) + 1 \leq n \leq S_k(\omega) + \bar{N}_k(\omega)$, $\hat{a}_n^k(\omega) \leq a_n(\omega)$;

FIG. 2. Sample paths of $\{\hat{a}_n^k\}$ and $\{a_n\}$.

(3) If $n = S_k(\omega) + \bar{N}_k(\omega)$, then $\hat{a}_n^k(\omega) = a_n(\omega)$.

The sequence $\{\hat{a}_n^k\}$ allows us to relate terms of the sequence $\{\theta_n\}$ for different customers in a BP without explicitly having to deal with projections. Whenever a projection occurs, say in computing θ_{n+1} , the value of a_n used in that iteration is “lost.” However, the value of \hat{h}_{n+1} may be added to the estimate for the next customer, and if so is retained. This occurs whenever the previous customer was not the first in a BP—hence there is no accumulation from one BP to the next. So within each BP the algorithm behaves like one in which there are no projections, but with updates occurring only when a prespecified criterion is met, namely, no projection is required, and with a derivative estimate that is an accumulation of the derivative estimates for each of the customers encompassed in the current interupdate interval. Note that in this interpretation the criterion for updating is $\theta_n - a_n \hat{h}_{n+1} \in D$, which is an \mathcal{F}_{n+1} -measurable event, and the random sequence $\{1 - p_{n+1}\}$ indicates precisely the occurrence of these events. We will have more to say about this in §4.

We are now ready to state and prove our lemma.

LEMMA 3.7. For each k and every m, n such that $S_k + 1 \leq m \leq n \leq S_k + \bar{N}_k$,

$$(12) \quad \theta_n - \hat{a}_n^k \hat{h}_{n+1} = \theta_m - \hat{a}_m^k \hat{h}_{m+1} - \sum_{i=m+1}^n \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i))$$

and

$$(13) \quad \theta_n - \theta_m = -\hat{a}_m^k \hat{h}_{m+1} - \sum_{i=m+1}^{l_n-1} \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i)),$$

where $l_n = \max\{l \in \mathbb{N} : l \leq n, p_l = 0\}$.

Proof. To show (12), it suffices to show that for $S_k + 1 < n \leq S_k + \bar{N}_k$,

$$\theta_n - \hat{a}_n^k \hat{h}_{n+1} = \theta_{n-1} - \hat{a}_{n-1}^k \hat{h}_n - \hat{a}_n^k (\hat{g}_{n+1} + C'(\theta_n)).$$

To see this, suppose first that $p_n = 0$. Then, $a_{n-1} = \hat{a}_{n-1}^k$, $\hat{h}_{n+1} = \hat{g}_{n+1} + C'(\theta_n)$, and $\theta_n = \theta_{n-1} - a_{n-1} \hat{h}_n$. Therefore, combining these three equations, we get

$$\theta_n - \hat{a}_n^k \hat{h}_{n+1} = \theta_{n-1} - \hat{a}_{n-1}^k \hat{h}_n - \hat{a}_n^k (\hat{g}_{n+1} + C'(\theta_n)).$$

Suppose now that $p_n = 1$. Then, $\hat{a}_n^k = \hat{a}_{n-1}^k$, $\hat{h}_{n+1} = \hat{g}_{n+1} + C'(\theta_n) + \hat{h}_n$, and $\theta_n = \theta_{n-1}$. Again, combining these three equations, we get

$$\theta_n - \hat{a}_n^k \hat{h}_{n+1} = \theta_{n-1} - \hat{a}_{n-1}^k \hat{h}_n - \hat{a}_n^k (\hat{g}_{n+1} + C'(\theta_n)).$$

To show (13), we rewrite (12) as

$$\begin{aligned} \theta_n - \theta_m &= -\hat{a}_m^k \hat{h}_{m+1} - \sum_{i=m+1}^n \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i)) + \hat{a}_n^k \hat{h}_{n+1} \\ &= -\hat{a}_m^k \hat{h}_{m+1} - \sum_{i=m+1}^{l_n-1} \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i)) + \hat{a}_n^k \hat{h}_{n+1} - \sum_{i=l_n}^n \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i)). \end{aligned}$$

Now, by definition of l_n , $p_{l_n+1} = \dots = p_n = 1$. Therefore, by (11), $\hat{a}_{l_n}^k = \dots = \hat{a}_n^k$, and by (7), $\hat{h}_{n+1} = \sum_{i=l_n}^n (\hat{g}_{i+1} + C'(\theta_i))$. Hence,

$$\hat{a}_n^k \hat{h}_{n+1} = \sum_{i=l_n}^n \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i))$$

and thus (13) holds, which completes the proof. \square

The following special case of (12) will be used frequently. When $m = S_k + 1$, $\hat{h}_{S_k+2} = \hat{g}_{S_k+2} + C'(\theta_{S_k+1})$, and (12) becomes

$$\theta_n - \hat{a}_n^k \hat{h}_{n+1} = \theta_m - \sum_{i=m}^n \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i)).$$

3.5. Proof of main result. The proof is lengthy, but the basic approach is very simple. The key feature is to first analyze the subsequence of $\{\theta_n\}$ at the start of the BPs. We show that this subsequence behaves like one obtained from an algorithm that updates at the start of BPs, driven by derivative estimates that are asymptotically unbiased. Then, using a standard convergence result for stochastic approximations, convergence of this subsequence is established. Finally, the convergence of the whole sequence $\{\theta_n\}$ is proved by showing that the θ_n values within each BP are “close” to the values at the beginning of the BPs. The main steps of the proof can be summarized as:

(M1) Showing that the subsequence of $\{\theta_n\}$ consisting of the values of θ at the start of every BP (denoted by $\{\tilde{\theta}_k\}$) satisfies the recursion

$$\tilde{\theta}_{k+1} = \tilde{\pi}_{k+1} [\tilde{\theta}_k - \tilde{b}_k \tilde{h}_{k+1}],$$

where $\{\tilde{b}_k\}$ is not summable almost surely, but is square summable, $\tilde{h}_{k+1} = dJ(\tilde{\theta}_k)/d\theta + \tilde{\varepsilon}_{k+1}$, where $\{\tilde{\varepsilon}_{k+1}\}$ asymptotically vanishes (in a sense to be defined later), and $\{\tilde{\pi}_{k+1}\}$ is the subsequence of $\{\pi_{n+1}\}$ taken at the start of each BP;

(M2) Showing that $\tilde{\theta}_k \rightarrow \theta^*$ almost surely; and

(M3) Showing that the whole sequence $\{\theta_n\}$ converges to θ^* almost surely.

To begin, let $\mathbf{X}^k = \{X_m^k : m \in \mathbb{N}\}$ and $\mathbf{A}^k = \{A_m^k : m \in \mathbb{N}\}$, and define a filtration $\{\tilde{\mathcal{F}}_k\}$ as follows: $\tilde{\mathcal{F}}_1 = \sigma(\theta_1)$ and for $k > 1$, $\tilde{\mathcal{F}}_k = \sigma(\tilde{\mathcal{F}}_{k-1}, \mathbf{X}^{k-1}, \mathbf{A}^{k-1})$. Then let $\tilde{\theta}_k = \theta_{S_k+1}$, $\tilde{a}_k = a_{S_k+1}$, and $\tilde{\pi}_k[\cdot] = \pi_{S_k+1}[\cdot]$, i.e.,

$$\tilde{\pi}_{k+1}[x] = \begin{cases} x, & \text{if } x \in D, \\ \theta_{S_k+\bar{N}_k}, & \text{otherwise.} \end{cases}$$

The sequences $\{\tilde{\theta}_k\}$, $\{\tilde{\pi}_k\}$ and $\{\tilde{a}_k\}$ are subsequences of $\{\theta_n\}$, $\{\pi_n\}$, and $\{a_n\}$, respectively, taken at the start of every BP. Note that $\tilde{\theta}_k$ and \tilde{a}_k are $\tilde{\mathcal{F}}_k$ -measurable.

Step (M1) of the proof is the subject of the following lemma.

LEMMA 3.8. *Suppose the assumptions of Theorem 3.2 are satisfied. Then there exist sequences $\{\tilde{h}_k\}$ and $\{\tilde{b}_k\}$ of random variables adapted to $\{\tilde{\mathcal{F}}_k\}$ such that*

(a) $\{\tilde{\theta}_k\}$ satisfies the recursion

$$(14) \quad \tilde{\theta}_{k+1} = \tilde{\pi}_{k+1}[\tilde{\theta}_k - \tilde{b}_k \tilde{h}_{k+1}].$$

(b) For all $k \in \mathbb{N}$, $\tilde{a}_k \leq \tilde{b}_k \leq B_N \tilde{a}_k$.

Moreover, if we let $\tilde{\varepsilon}_{k+1} = \tilde{h}_{k+1} - dJ(\tilde{\theta}_k)/d\theta$, then

(c) $|E_{\tilde{\mathcal{F}}_k}(\tilde{\varepsilon}_{k+1})| \leq B \tilde{a}_k + B_d \tilde{a}_k^d$ for constants $B, B_d < \infty$, and

(d) $E_{\tilde{\mathcal{F}}_k}(\tilde{\varepsilon}_{k+1}^2) \leq \sigma^2$ for a constant $\sigma^2 < \infty$.

Proof. Without loss of generality, we may let $k = 1$ (this greatly simplifies the notation). Note that $S_1 = 0$, $\tilde{\theta}_1 = \theta_1$, $\tilde{a}_1 = a_1$, and $\tilde{\mathcal{F}}_1 = \mathcal{F}_1$. To further simplify the notation, let $N = N_1(\theta_1)$, $\bar{N} = \bar{N}_1$, $\bar{\phi}_n = \bar{\phi}_n^1$, and $K_n = K_n^1$.

Define the sequence $\{\theta_n^1\}$ by the following: For $1 \leq n \leq \bar{N}$, $\theta_n^1 = \theta_n$, and for $n > \bar{N}$, $\theta_n^1 = \theta_{\bar{N}}$. So, $\{\theta_n^1\}$ is simply an extension of $\{\theta_n : 1 \leq n \leq \bar{N}\}$, the parameter sequence within the first BP, obtained by repeating the last value $\theta_{\bar{N}}$ indefinitely.

We first prove (a) and (b). Now, by using the definitions of $\tilde{\theta}_2$ and $\tilde{\pi}_2$, and property (3) of $\{\hat{a}_n^1\}$, we have

$$\tilde{\theta}_2 = \theta_{\bar{N}+1} = \pi_{\bar{N}+1}[\theta_{\bar{N}} - a_{\bar{N}} \hat{h}_{\bar{N}+1}] = \tilde{\pi}_2[\theta_{\bar{N}} - \hat{a}_{\bar{N}}^1 \hat{h}_{\bar{N}+1}]$$

and then using Lemma 3.7 with $m = 1$ and $n = \bar{N}$, we obtain

$$\tilde{\theta}_2 = \tilde{\pi}_2 \left[\theta_1 - \sum_{i=1}^{\bar{N}} \hat{a}_i^1 (\hat{g}_{i+1} + C'(\theta_i)) \right].$$

Now, by (6),

$$\begin{aligned} & \sum_{i=1}^{\bar{N}} \hat{a}_i^1 (\hat{g}_{i+1} + C'(\theta_i)) \\ &= \sum_{i=1}^{\bar{N}} \hat{a}_i^1 \left(\sum_{j=1}^i \phi_j^1(\theta_j^1) + C'(\theta_i^1) \right) \\ &= a_1 E_{\theta_1}(N) \sum_{i=1}^{\bar{N}} \frac{\hat{a}_i^1}{a_1} \left(\sum_{j=1}^i \phi_j^1(\theta_j^1) + C'(\theta_i^1) \right) \frac{1}{E_{\theta_1}(N)} \\ &= \tilde{a}_1 E_{\theta_1}(N) \left(\frac{1}{E_{\theta_1}(N)} \sum_{i=1}^{\bar{N}} \frac{\hat{a}_i^1}{a_1} \sum_{j=1}^i \phi_j^1(\theta_j^1) + \frac{1}{E_{\theta_1}(N)} \sum_{i=1}^{\bar{N}} \frac{\hat{a}_i^1}{a_1} C'(\theta_i^1) \right) \\ &= \tilde{a}_1 E_{\theta_1}(N) (\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3), \end{aligned}$$

where \tilde{c}_1 , \tilde{c}_2 , and \tilde{c}_3 are defined as follows:

$$\tilde{c}_1 = \frac{1}{E_{\theta_1}(N)} \sum_{i=1}^{\bar{N}} \frac{\hat{a}_i^1}{a_1} \sum_{j=1}^i \phi_j^1(\theta_j^1),$$

$$\tilde{c}_2 = \frac{1}{E_{\theta_1}(N)} \sum_{i=1}^N \frac{\hat{a}_i^1}{a_1} C'(\theta_i^1),$$

$$\tilde{c}_3 = \begin{cases} \frac{1}{E_{\theta_1}(N)} \sum_{i=\bar{N}+1}^{\bar{N}} \frac{\hat{a}_i^1}{a_1} \left(\sum_{j=1}^i \phi_j^1(\theta_j^1) + C'(\theta_i^1) \right), & \text{if } \bar{N} > N, \\ 0, & \text{if } \bar{N} = N, \\ -\frac{1}{E_{\theta_1}(N)} \sum_{i=\bar{N}+1}^N \frac{\hat{a}_i^1}{a_1} \left(\sum_{j=1}^i \phi_j^1(\theta_j^1) + C'(\theta_i^1) \right), & \text{if } \bar{N} < N. \end{cases}$$

Let $\tilde{h}_2 = \tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3$, and $\tilde{b}_1 = E_{\theta_1}(N)\tilde{a}_1$. Then, we may write $\tilde{\theta}_2 = \tilde{\pi}_2[\tilde{\theta}_1 - \tilde{b}_1\tilde{h}_2]$. It is easy to check that \tilde{h}_2 is $\tilde{\mathcal{F}}_2$ -measurable. Since $E_{\theta_1}(N)$ and \tilde{a}_1 are $\tilde{\mathcal{F}}_1$ -measurable, then so is \tilde{b}_1 . This proves (a).

It is clear that $E_{\theta_1}(N) \geq 1$. Also, using Lemma 3.3, $E_{\theta_1}(N) \leq E_{\theta_1}(N_1(\theta_{\max})) = E(N_1(\theta_{\max})) \leq B_N$. Therefore, $\tilde{a}_1 \leq \tilde{b}_1 \leq B_N\tilde{a}_1$. This proves (b).

We now show (c). We can write $\tilde{\varepsilon}_2$ as

$$(15) \quad \begin{aligned} \tilde{\varepsilon}_2 &= \tilde{h}_2 - (T'(\theta_1) + C'(\theta_1)) \\ &= (\tilde{c}_1 - T'(\theta_1)) + (\tilde{c}_2 - C'(\theta_1)) + \tilde{c}_3 \end{aligned}$$

and hence

$$(16) \quad |E_{\mathcal{F}_1}(\tilde{\varepsilon}_2)| \leq |E_{\mathcal{F}_1}(\tilde{c}_1 - T'(\theta_1))| + |E_{\mathcal{F}_1}(\tilde{c}_2 - C'(\theta_1))| + |E_{\mathcal{F}_1}(\tilde{c}_3)|.$$

Since $\tilde{\mathcal{F}}_1 = \mathcal{F}_1$, to show that $|E_{\tilde{\mathcal{F}}_1}(\tilde{\varepsilon}_2)| \leq B\tilde{a}_1 + B_d\tilde{a}_1^d$, it suffices to show that $|E_{\mathcal{F}_1}(\tilde{c}_1 - T'(\theta_1))| \leq B_1a_1$, $|E_{\mathcal{F}_1}(\tilde{c}_2 - C'(\theta_1))| \leq B_2a_1$, and $|E_{\mathcal{F}_1}(\tilde{c}_3)| \leq B_3a_1^d$, where $B_1, B_2, B_3 < \infty$. This will proceed in several steps.

Step 1. Show that $|E_{\mathcal{F}_1}(\tilde{c}_1 - T'(\theta_1))| \leq B_1a_1$.

Now, since θ_1 is \mathcal{F}_1 -measurable, then $E_{\mathcal{F}_1}(\tilde{c}_1 - T'(\theta_1)) = E_{\mathcal{F}_1}(\tilde{c}_1) - T'(\theta_1)$. By Lemma 2.1 and the fact that θ_1 is independent of \mathbf{A}^1 and \mathbf{X}^1 we may write

$$\begin{aligned} T'(\theta_1) &= \frac{1}{E_{\theta_1}(N)} E_{\theta_1} \left(\sum_{i=1}^N \sum_{j=1}^i \phi_j^1(\theta_1) \right) \\ &= \frac{1}{E_{\theta_1}(N)} E_{\mathcal{F}_1} \left(\sum_{i=1}^N \sum_{j=1}^i \phi_j^1(\theta_1) \right), \end{aligned}$$

where the last line holds because the random variable inside the conditional expectation operator depends on \mathcal{F}_1 only through θ_1 . Thus, we may write

$$\begin{aligned} E_{\mathcal{F}_1}(\tilde{c}_1 - T'(\theta_1)) &= \frac{1}{E_{\theta_1}(N)} E_{\mathcal{F}_1} \left(\sum_{i=1}^N \frac{\hat{a}_i^1}{a_1} \sum_{j=1}^i \phi_j^1(\theta_j^1) \right) - \frac{1}{E_{\theta_1}(N)} E_{\mathcal{F}_1} \left(\sum_{i=1}^N \sum_{j=1}^i \phi_j^1(\theta_1) \right) \\ &= \frac{1}{E_{\theta_1}(N)} E_{\mathcal{F}_1} \left(\sum_{i=1}^N \left(\frac{\hat{a}_i^1}{a_1} - 1 \right) \sum_{j=1}^i \phi_j^1(\theta_j^1) + \sum_{i=1}^N \sum_{j=1}^i \phi_j^1(\theta_j^1) - \phi_j^1(\theta_1) \right), \end{aligned}$$

and, since $E_{\theta_1}(N) \geq 1$,

$$(17) \quad |E_{\mathcal{F}_1}(\tilde{c}_1 - T'(\theta_1))| \leq E_{\mathcal{F}_1} \left(\sum_{i=1}^N \left| \frac{\hat{a}_i^1}{a_1} - 1 \right| \sum_{j=1}^i |\phi_j^1(\theta_j^1)| + \sum_{i=1}^N \sum_{j=1}^i |\phi_j^1(\theta_j^1) - \phi_j^1(\theta_1)| \right).$$

Now, using the properties of the sequence $\{\hat{a}_i^1\}$, and assumptions (G2) and (G4),

$$\begin{aligned}
 \left| \frac{\hat{a}_i^1}{a_1} - 1 \right| &= \hat{a}_i^1 \left(\frac{1}{\hat{a}_i^1} - \frac{1}{a_1} \right) \\
 &\leq a_1 \left(\frac{1}{a_{\bar{N}}} - \frac{1}{a_1} \right) \\
 (18) \quad &= a_1 \left(\frac{1}{a_{\bar{N}}} - \frac{1}{a_{\bar{N}-1}} + \frac{1}{a_{\bar{N}-1}} - \frac{1}{a_{\bar{N}-2}} + \cdots - \frac{1}{a_1} \right) \\
 &\leq a_1 \bar{N} B_a.
 \end{aligned}$$

Also, by assumption (S2),

$$(19) \quad \sum_{j=1}^i |\phi_j^1(\theta_j^1)| \leq \sum_{j=1}^i \bar{\phi}_j \leq \sum_{j=1}^N \bar{\phi}_j.$$

For each k , let $N_{\max}^k \triangleq N_k(\theta_{\max})$. By Lemma 3.3, $E(N_{\max}^k)^r \leq B_N$ for all $r \leq p$. It is clear by the construction of the BP that $\bar{N}_k \leq N_{\max}^k$. For brevity, write $N_{\max} = N_{\max}^1$.

Combining (18) and (19) yields

$$\begin{aligned}
 E_{\mathcal{F}_1} \left(\sum_{i=1}^N \left| \frac{\hat{a}_i^1}{a_1} - 1 \right| \sum_{j=1}^i |\phi_j^1(\theta_j^1)| \right) &\leq E_{\mathcal{F}_1} \left(a_1 N \bar{N} B_a \sum_{j=1}^N \bar{\phi}_j \right) \\
 &\leq a_1 B_a E_{\mathcal{F}_1} \left(N_{\max}^2 \sum_{j=1}^{N_{\max}} \bar{\phi}_j \right).
 \end{aligned}$$

Using Schwarz's inequality, the fact that N_{\max} is independent of \mathcal{F}_1 , and Lemmas 3.3 and 3.4, we then obtain

$$\begin{aligned}
 E_{\mathcal{F}_1} \left(\sum_{i=1}^N \left| \frac{\hat{a}_i^1}{a_1} - 1 \right| \sum_{j=1}^i |\phi_j^1(\theta_j^1)| \right) &\leq a_1 B_a \sqrt{E(N_{\max}^4)} \sqrt{E \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j \right)^2} \\
 (20) \quad &\leq a_1 B_a \sqrt{B_N} \sqrt{B_{\phi}}.
 \end{aligned}$$

Next, by assumption (S3),

$$(21) \quad \sum_{j=1}^i |\phi_j^1(\theta_j^1) - \phi_j^1(\theta_1)| \leq \sum_{j=1}^i K_j |\theta_j^1 - \theta_1|.$$

By Lemma 3.7,

$$\theta_j^1 - \theta_1 = - \sum_{i=1}^{k_j} \hat{a}_i^1 (\hat{g}_{i+1} + C'(\theta_i^1)),$$

where $k_j = \max\{k \in \mathbb{N} : k \leq j, p_k = 0\}$, and using (S2) and the fact that C' is bounded we obtain

$$|\theta_j^1 - \theta_1| \leq a_1 \sum_{i=1}^{\bar{N}} (|\hat{g}_{i+1}| + |C'(\theta_i^1)|)$$

$$\begin{aligned}
 (22) \quad & \leq a_1 \sum_{i=1}^{\bar{N}} \left(\sum_{k=1}^i |\phi_k^1(\theta_k^1)| + |C'(\theta_i^1)| \right) \\
 & \leq a_1 \bar{N} \left(\sum_{k=1}^{\bar{N}} \bar{\phi}_k + C'_M \right).
 \end{aligned}$$

Thus, from (21) and (22) we have

$$\begin{aligned}
 \sum_{i=1}^N \sum_{j=1}^i |\phi_j^1(\theta_j^1) - \phi_j^1(\theta_1)| & \leq \sum_{i=1}^N \sum_{j=1}^N a_1 K_j \bar{N} \left(\sum_{k=1}^{\bar{N}} \bar{\phi}_k + C'_M \right) \\
 & \leq a_1 N_{\max}^2 \left(\sum_{j=1}^N K_j \right) \left(\sum_{k=1}^{N_{\max}} \bar{\phi}_k + C'_M \right).
 \end{aligned}$$

Taking conditional expectations with respect to \mathcal{F}_1 , using Schwarz's inequality, (10), and Lemmas 3.3 and 3.5, yields

$$\begin{aligned}
 (23) \quad E_{\mathcal{F}_1} \left(\sum_{i=1}^N \sum_{j=1}^i |\phi_j^1(\theta_j^1) - \phi_j^1(\theta_1)| \right) & \leq a_1 E_{\mathcal{F}_1} \left(N_{\max}^2 \left(\sum_{j=1}^{N_{\max}} K_j \right) \left(\sum_{k=1}^{N_{\max}} \bar{\phi}_k + C'_M \right) \right) \\
 & \leq a_1 \sqrt{B_K} \sqrt{\sqrt{B_N} \sqrt{B'_\phi}}.
 \end{aligned}$$

Finally, combining (20) and (23) with (17), we get

$$|E_{\mathcal{F}_1}(\tilde{c}_1 - T'(\theta_1))| \leq a_1 B_a \sqrt{B_N} \sqrt{B_\phi} + a_1 \sqrt{B_K} \sqrt{\sqrt{B_N} \sqrt{B'_\phi}} \leq B_1 a_1,$$

where $B_1 = B_a \sqrt{B_N} \sqrt{B_\phi} + \sqrt{B_K} \sqrt{\sqrt{B_N} \sqrt{B'_\phi}} < \infty$.

Step 2. Show that $|E_{\mathcal{F}_1}(\tilde{c}_2 - C'(\theta_1))| \leq B_2 a_1$.

Now,

$$\begin{aligned}
 E_{\mathcal{F}_1}(\tilde{c}_2 - C'(\theta_1)) & = \frac{1}{E_{\theta_1}(N)} E_{\mathcal{F}_1} \left(\sum_{i=1}^N \frac{\hat{a}_i^1}{a_1} C'(\theta_i^1) - \sum_{i=1}^N C'(\theta_1) \right) \\
 & = \frac{1}{E_{\theta_1}(N)} E_{\mathcal{F}_1} \left(\sum_{i=1}^N \left(\frac{\hat{a}_i^1}{a_1} - 1 \right) C'(\theta_i^1) + \sum_{i=1}^N C'(\theta_i^1) - C'(\theta_1) \right),
 \end{aligned}$$

and therefore

$$(24) \quad |E_{\mathcal{F}_1}(\tilde{c}_2 - C'(\theta_1))| \leq E_{\mathcal{F}_1} \left(\sum_{i=1}^N \left| \frac{\hat{a}_i^1}{a_1} - 1 \right| |C'(\theta_i^1)| + \sum_{i=1}^N |C'(\theta_i^1) - C'(\theta_1)| \right).$$

By (18), the fact that C' is bounded, and Lemma 3.3, we have that

$$\begin{aligned}
 (25) \quad E_{\mathcal{F}_1} \left(\sum_{i=1}^N \left| \frac{\hat{a}_i^1}{a_1} - 1 \right| |C'(\theta_i^1)| \right) & \leq E_{\mathcal{F}_1} \left(\sum_{i=1}^N a_1 \bar{N} B_a C'_M \right) \\
 & \leq a_1 C'_M B_a E(N_{\max}^2) \\
 & \leq a_1 C'_M B_a B_N.
 \end{aligned}$$

Also, by assumption (C2), and (22),

$$\begin{aligned} \sum_{i=1}^N |C'(\theta_i^1) - C'(\theta_1)| &\leq \sum_{i=1}^N C''_M |\theta_i^1 - \theta_1| \\ &\leq \sum_{i=1}^N C''_M a_1 \bar{N} \left(\sum_{j=1}^{\bar{N}} \bar{\phi}_j + C'_M \right) \\ &\leq a_1 C''_M N_{\max}^2 \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j + C'_M \right). \end{aligned}$$

Taking conditional expectations with respect to \mathcal{F}_1 , using Schwarz's inequality, (10), and Lemma 3.3 yields

$$\begin{aligned} E_{\mathcal{F}_1} \left(\sum_{i=1}^N |C'(\theta_i^1) - C'(\theta_1)| \right) &\leq a_1 C''_M E_{\mathcal{F}_1} \left(N_{\max}^2 \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j + C'_M \right) \right) \\ (26) \qquad \qquad \qquad &\leq a_1 C''_M \sqrt{B_N} \sqrt{B'_\phi}. \end{aligned}$$

Finally, combining (26) and (26) with (24), we get

$$|E_{\mathcal{F}_1}(\tilde{c}_2 - C'(\theta_1))| \leq a_1 C'_M B_a B_N + a_1 C''_M \sqrt{B_N} \sqrt{B'_\phi} \leq B_2 a_1,$$

where $B_2 = C'_M B_a B_N + C''_M \sqrt{B_N} \sqrt{B'_\phi} < \infty$.

Step 3. Show that $|E_{\mathcal{F}_1}(\tilde{c}_3)| \leq B_3 a_1^d$.

Let 1_A denote the indicator function of the set A . Then, by definition of \tilde{c}_3 , we can write

$$\begin{aligned} \tilde{c}_3 &= 1_{\{\bar{N} > N\}} \frac{1}{E_{\theta_1}(N)} \sum_{i=N+1}^{\bar{N}} \frac{\hat{a}_i^1}{a_1} \left(\sum_{j=1}^i \phi_j^1(\theta_j^1) + C'(\theta_i^1) \right) \\ &\quad - 1_{\{\bar{N} < N\}} \frac{1}{E_{\theta_1}(N)} \sum_{i=\bar{N}+1}^N \frac{\hat{a}_i^1}{a_1} \left(\sum_{j=1}^i \phi_j^1(\theta_j^1) + C'(\theta_i^1) \right), \end{aligned}$$

and thus

$$\begin{aligned} |\tilde{c}_3| &\leq (1_{\{\bar{N} > N\}} + 1_{\{\bar{N} < N\}}) \frac{1}{E_{\theta_1}(N)} \sum_{i=1}^{N_{\max}} \left(\sum_{j=1}^i \bar{\phi}_j + C'_M \right) \\ (27) \qquad \qquad \qquad &\leq 1_{\{\bar{N} \neq N\}} N_{\max} \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j + C'_M \right). \end{aligned}$$

Let $p_1 = (1+d)/2d$ and $q_1 = (1+d)/(1-d)$. Note that p_1 and q_1 are conjugate exponents, i.e., $1/p_1 + 1/q_1 = 1$ and $p_1, q_1 > 1$. Then, by Hölder's inequality, and

Schwarz's inequality,

$$\begin{aligned} |E_{\mathcal{F}_1}(\tilde{c}_3)| &\leq E_{\mathcal{F}_1}(|\tilde{c}_3|) \\ &\leq (E_{\mathcal{F}_1}(1_{\{\bar{N} \neq N\}})^{p_1})^{1/p_1} \left(E_{\mathcal{F}_1} \left(N_{\max} \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j + C'_M \right) \right)^{q_1} \right)^{1/q_1} \\ &\leq (E_{\mathcal{F}_1}(1_{\{\bar{N} \neq N\}}))^{1/p_1} \hat{B}_1, \end{aligned}$$

where the last line follows from (10) and Lemma 3.3 with $\hat{B}_1 = (\sqrt{B_N} \sqrt{B'_\phi})^{1/q_1} < \infty$.

Now, $E_{\mathcal{F}_1}(1_{\{\bar{N} \neq N\}}) = P_{\mathcal{F}_1}\{\bar{N} \neq N\} = P_{\mathcal{F}_1}\{\bar{N} > N\} + P_{\mathcal{F}_1}\{\bar{N} < N\}$. Therefore, to show that $|E_{\mathcal{F}_1}(\tilde{c}_3)| \leq B_3 a_1^d$, it suffices to show that $P_{\mathcal{F}_1}\{\bar{N} > N\} \leq \hat{B}_2 a_1^{(1+d)/2}$ and $P_{\mathcal{F}_1}\{\bar{N} < N\} \leq \hat{B}_3 a_1^{(1+d)/2}$ where $\hat{B}_2, \hat{B}_3 < \infty$.

Step 3(a). Show that $P_{\mathcal{F}_1}\{\bar{N} > N\} \leq \hat{B}_2 a_1^{(1+d)/2}$.

By definition, N is the smallest positive integer satisfying

$$\sum_{i=1}^N X_i^1(\theta_1) < \sum_{i=1}^N A_i^1.$$

The positive random variable $I_1 = \sum_{i=1}^N A_i^1 - \sum_{i=1}^N X_i^1(\theta_1)$ is the duration of the idle period following the first BP assuming the control parameter is given by θ_1 throughout the BP. Similarly, \bar{N} is the smallest positive integer satisfying

$$\sum_{i=1}^{\bar{N}} X_i^1(\theta_1^1) < \sum_{i=1}^{\bar{N}} A_i^1.$$

The positive random variable $\bar{I}_1 = \sum_{i=1}^{\bar{N}} A_i^1 - \sum_{i=1}^{\bar{N}} X_i^1(\theta_1^1)$ is the duration of the *actual* idle period following the first actual BP. We have that

$$\begin{aligned} \bar{N} > N &\Rightarrow \sum_{i=1}^N X_i^1(\theta_1^1) \geq \sum_{i=1}^N A_i^1 \\ &\Rightarrow \sum_{i=1}^N (X_i^1(\theta_1^1) - X_i^1(\theta_1)) \geq \sum_{i=1}^N (A_i^1 - X_i^1(\theta_1)) = I_1 \\ &\Rightarrow \sum_{i=1}^N |X_i^1(\theta_1^1) - X_i^1(\theta_1)| \geq I_1. \end{aligned}$$

Thus,

$$P_{\mathcal{F}_1}\{\bar{N} > N\} \leq P_{\mathcal{F}_1} \left\{ \sum_{i=1}^N |X_i^1(\theta_1^1) - X_i^1(\theta_1)| \geq I_1 \right\}.$$

Let E_I be the event $\left\{ \sum_{i=1}^N |X_i^1(\theta_1^1) - X_i^1(\theta_1)| \geq I_1 \right\}$. It is clear that

$$1_{E_I} \leq \frac{1}{I_1^{(1+d)/2}} \left(\sum_{i=1}^N |X_i^1(\theta_1^1) - X_i^1(\theta_1)| \right)^{(1+d)/2}.$$

Therefore,

$$P_{\mathcal{F}_1}\{\bar{N} > N\} \leq E_{\mathcal{F}_1} \left(\frac{1}{I_1^{(1+d)/2}} \left(\sum_{i=1}^N |X_i^1(\theta_i^1) - X_i^1(\theta_1)| \right)^{(1+d)/2} \right).$$

Let $p_2 = (3+d)/(2(1+d))$ and $q_2 = (3+d)/(1-d)$. Then p_2 and q_2 are conjugate exponents, and by Hölder's inequality,

$$P_{\mathcal{F}_1}\{\bar{N} > N\} \leq \left(E_{\mathcal{F}_1} \left(\frac{1}{I_1^{(1+d)/2}} \right)^{p_2} \right)^{1/p_2} \left(E_{\mathcal{F}_1} \left(\sum_{i=1}^N |X_i^1(\theta_i^1) - X_i^1(\theta_1)| \right)^{q_2(1+d)/2} \right)^{1/q_2}.$$

Let $p_3 = p_2(1+d)/2 = (3+d)/4 < 1$, and $q_3 = q_2(1+d)/2 = (1+d)(3+d)/(2(1-d))$. Then, we have that

$$(28) \quad P_{\mathcal{F}_1}\{\bar{N} > N\} \leq \left(E_{\mathcal{F}_1} \left(\frac{1}{I_1^{p_3}} \right) \right)^{1/p_2} \left(E_{\mathcal{F}_1} \left(\sum_{i=1}^N |X_i^1(\theta_i^1) - X_i^1(\theta_1)| \right)^{q_3} \right)^{1/q_2}.$$

Now, by the Mean Value Theorem, assumption (S2), and (22),

$$\begin{aligned} \sum_{i=1}^N |X_i^1(\theta_i^1) - X_i^1(\theta_1)| &\leq \sum_{i=1}^N \bar{\phi}_i |\theta_i^1 - \theta_1| \\ &\leq \sum_{i=1}^N \bar{\phi}_i a_1 \bar{N} \left(\sum_{j=1}^{\bar{N}} \bar{\phi}_j + C'_M \right) \\ &\leq a_1 N_{\max} \left(\sum_{i=1}^N \bar{\phi}_i \right) \left(\sum_{j=1}^{\bar{N}} \bar{\phi}_j + C'_M \right), \end{aligned}$$

and thus using Schwarz's inequality and our known bounds,

$$\begin{aligned} (29) \quad &E_{\mathcal{F}_1} \left(\sum_{i=1}^N |X_i^1(\theta_i^1) - X_i^1(\theta_1)| \right)^{q_3} \\ &\leq E_{\mathcal{F}_1} \left(a_1 N_{\max} \left(\sum_{i=1}^N \bar{\phi}_i \right) \left(\sum_{j=1}^{\bar{N}} \bar{\phi}_j + C'_M \right) \right)^{q_3} \\ &\leq a_1^{q_3} \sqrt{\sqrt{B_N} \sqrt{B_\phi} \sqrt{B'_\phi}}. \end{aligned}$$

Combining (29) with (28), we get

$$P_{\mathcal{F}_1}\{\bar{N} > N\} \leq \left(E_{\theta_1} \left(\frac{1}{I_1^{p_3}} \right) \right)^{1/p_2} \left(a_1^{q_3} \sqrt{\sqrt{B_N} \sqrt{B_\phi} \sqrt{B'_\phi}} \right)^{1/q_2}.$$

Since $p_3 < 1$, then by Lemma 3.6, we have

$$P_{\mathcal{F}_1}\{\bar{N} > N\} \leq a_1^{(1+d)/2} \hat{B}_2,$$

where $\hat{B}_2 = (B_I)^{1/p_2} \left(\sqrt{\sqrt{B_N} \sqrt{B_\phi} \sqrt{B'_\phi}} \right)^{1/q_2} < \infty$.

Step 3(b). Show that $P_{\mathcal{F}_1}\{\bar{N} < N\} \leq \hat{B}_3 a_1^{(1+d)/2}$.

As before, we may write

$$\begin{aligned} \bar{N} < N &\Rightarrow \sum_{i=1}^{\bar{N}} X_i^1(\theta_1) \geq \sum_{i=1}^{\bar{N}} A_i^1 \\ &\Rightarrow \sum_{i=1}^{\bar{N}} X_i^1(\theta_1) - X_i^1(\theta_i^1) \geq \sum_{i=1}^{\bar{N}} A_i^1 - X_i^1(\theta_i^1) = \bar{I}_1 \\ &\Rightarrow \sum_{i=1}^{\bar{N}} |X_i^1(\theta_1) - X_i^1(\theta_i^1)| \geq \bar{I}_1. \end{aligned}$$

Thus,

$$P_{\mathcal{F}_1}\{\bar{N} < N\} \leq P_{\mathcal{F}_1}\left\{\sum_{i=1}^{\bar{N}} |X_i^1(\theta_1) - X_i^1(\theta_i^1)| \geq \bar{I}_1\right\}.$$

Proceeding in a similar fashion as in Step 3(a) (replacing all occurrences of N with \bar{N} , and I_1 with \bar{I}_1), we get that $P_{\mathcal{F}_1}\{\bar{N} < N\} \leq \hat{B}_3 a_1^{(1+d)/2}$ where $\hat{B}_3 < \infty$, which proves the result of Step 3(b).

Using Steps 3(a) and 3(b), we may write

$$\begin{aligned} |E_{\mathcal{F}_1}(\tilde{c}_3)| &\leq \left(a_1^{(1+d)/2}(\hat{B}_2 + \hat{B}_3)\right)^{\frac{1}{p_1}} \hat{B}_1 \\ (30) \quad &= a_1^{(1+d)/(2p_1)}(\hat{B}_2 + \hat{B}_3)^{\frac{1}{p_1}} \hat{B}_1 \\ &= a_1^d B_3, \end{aligned}$$

where $B_3 = (\hat{B}_2 + \hat{B}_3)^{1/p_1} \hat{B}_1 < \infty$. This completes Step 3.

Step 4. Show that $|E_{\tilde{\mathcal{F}}_n}(\tilde{\varepsilon}_{n+1})| \leq B\tilde{a}_n + B_d \tilde{a}_n^d$.

Using (16) and Steps 1–3, we have that

$$|E_{\tilde{\mathcal{F}}_n}(\tilde{\varepsilon}_2)| \leq B_1 a_1 + B_2 a_1 + B_3 a_1^d = B\tilde{a}_1 + B_d \tilde{a}_1^d,$$

where $B = (B_1 + B_2) < \infty$ and $B_d = B_3 < \infty$. Thus we have shown (c).

We now show (d). From (15),

$$\begin{aligned} \tilde{\varepsilon}_2^2 &= (\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3 - T'(\theta_1) - C'(\theta_1))^2 \\ (31) \quad &\leq (|\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3| + T'_M + C'_M)^2 \\ &\leq 2(\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3)^2 + 2(T'_M + C'_M)^2, \end{aligned}$$

where $T'_M \triangleq \sup_{\theta \in D} |T'(\theta)|$. By Lemma 2.1, T' is continuous, and since D is compact, $T'_M < \infty$. Taking conditional expectations with respect to \mathcal{F}_1 ,

$$(32) \quad E_{\mathcal{F}_1}(\tilde{\varepsilon}_2^2) \leq 2E_{\mathcal{F}_1}(\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3)^2 + 2(T'_M + C'_M)^2.$$

Now,

$$\begin{aligned} (33) \quad E_{\mathcal{F}_1}(\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3)^2 &\leq 4E_{\mathcal{F}_1}(\tilde{c}_1^2 + \tilde{c}_2^2 + \tilde{c}_3^2) \\ &= 4E_{\mathcal{F}_1}(\tilde{c}_1^2) + 4E_{\mathcal{F}_1}(\tilde{c}_2^2) + 4E_{\mathcal{F}_1}(\tilde{c}_3^2). \end{aligned}$$

Therefore, it will suffice to show that $E_{\mathcal{F}_1}(\tilde{c}_1^2) \leq \sigma_1^2$, $E_{\mathcal{F}_1}(\tilde{c}_2^2) \leq \sigma_2^2$, and $E_{\mathcal{F}_1}(\tilde{c}_3^2) \leq \sigma_3^2$, where $\sigma_1^2, \sigma_2^2, \sigma_3^2 < \infty$.

First,

$$\tilde{c}_1^2 = \frac{1}{(E_{\theta_1}(N))^2} \left(\sum_{i=1}^N \frac{\hat{a}_i^1}{a_1} \sum_{j=1}^i \phi_j^1(\theta_j^1) \right)^2 \leq \left(N_{\max} \sum_{j=1}^{N_{\max}} \bar{\phi}_j \right)^2.$$

Therefore, using Schwarz's inequality and Lemmas 3.3 and 3.4,

$$(34) \quad E_{\mathcal{F}_1}(\tilde{c}_1^2) \leq E_{\mathcal{F}_1} \left(N_{\max}^2 \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j \right)^2 \right) \leq \sigma_1^2,$$

where $\sigma_1^2 = \sqrt{B_N} \sqrt{B_\phi} < \infty$.

Next,

$$\tilde{c}_2^2 = \frac{1}{(E_{\theta_1}(N))^2} \left(\sum_{i=1}^N C'(\theta_i^1) \right)^2 \leq N^2 (C'_M)^2.$$

Hence by Lemma 3.3, $E_{\mathcal{F}_1}(\tilde{c}_2^2) \leq E_{\theta_1}(N^2 (C'_M)^2) \leq \sigma_2^2$, where $\sigma_2^2 = (C'_M)^2 B_N < \infty$.

Finally, by (27), we have

$$\tilde{c}_3^2 \leq N_{\max}^2 \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j + C'_M \right)^2,$$

and hence using Schwarz's inequality, (10), and Lemma 3.3,

$$(35) \quad E_{\mathcal{F}_1}(\tilde{c}_3^2) \leq E_{\mathcal{F}_1} \left(N_{\max}^2 \left(\sum_{j=1}^{N_{\max}} \bar{\phi}_j + C'_M \right)^2 \right) \leq \sigma_3^2,$$

where $\sigma_3^2 = \sqrt{B_N} \sqrt{B'_\phi} < \infty$.

Let $\hat{\sigma}^2 = 4\sigma_1^2 + 4\sigma_2^2 + 4\sigma_3^2 < \infty$. Then, from (33), $E_{\mathcal{F}_1}(\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3)^2 \leq \hat{\sigma}^2$. Combining the above with (32), we get $E_{\mathcal{F}_1}(\tilde{\varepsilon}_2^2) \leq \sigma^2$, where $\sigma^2 = 2\hat{\sigma}^2 + 2(T'_M + C'_M)^2 < \infty$. Note that σ^2 is independent of θ_1 , and of the fact that $k = 1$. \square

Having proven step (M1), step (M2) can be shown with the aid of the following lemma.

LEMMA 3.9. *Suppose conditions (P1)–(P5), (S3) and (C1)–(C3) hold. Let $\{\tilde{\theta}_k\}$ satisfy (14). Suppose $\{\tilde{b}_k\}$ satisfies:*

(A1) $\sum_{k=1}^{\infty} \tilde{b}_k = \infty$ almost surely;

(A2) $\sum_{k=1}^{\infty} \tilde{b}_k^2 < \infty$ almost surely;

and $\{\tilde{\varepsilon}_k\}$ satisfies:

(E1) $\sum_{k=1}^{\infty} \tilde{b}_k |E_{\tilde{\mathcal{F}}_k}(\tilde{\varepsilon}_{k+1})| < \infty$ almost surely;

(E2) For all k , $E_{\tilde{\mathcal{F}}_k}(\tilde{\varepsilon}_{k+1}^2) \leq \sigma^2$, where σ^2 is a finite constant.

Then, $\tilde{\theta}_k \rightarrow \theta^*$ as $k \rightarrow \infty$ with probability one.

Proof. The proof is a variation of a standard result on the convergence of stochastic approximation algorithms. A detailed proof based on martingale convergence arguments is available in [8]. \square

Step (M2) is the result of the following lemma.

LEMMA 3.10. *Suppose the assumptions of Theorem 3.2 hold. Then, $\tilde{\theta}_k \rightarrow \theta^*$ as $k \rightarrow \infty$ with probability one.*

Proof. We prove this result by appealing to Lemma 3.9. To do so, we need only verify that $\{\tilde{b}_k\}$ satisfies conditions (A1) and (A2), and that $\{\tilde{\varepsilon}_k\}$ satisfies (E1) and (E2), since all the other assumptions of Lemma 3.9 already hold.

We first claim that for each k , $\tilde{a}_k \geq A_l a_k$ where A_l is a random variable such that $A_l > 0$ almost surely. To see this, we write

$$\frac{\tilde{a}_k}{a_k} = \frac{a_{S_{k+1}}}{a_k} \geq \frac{a_{S_{n+1}}}{a_k}.$$

Then by assumption (G3),

$$(36) \quad \frac{\tilde{a}_k}{a_k} \geq \frac{A}{\bar{A}} \left(\frac{S_{n+1}}{k} \right)^{-\delta}.$$

Now,

$$\frac{S_{n+1}}{k} = \frac{1}{k} \sum_{i=1}^k \tilde{N}_i \leq \frac{1}{k} \sum_{i=1}^k N_{\max}^i.$$

Since $\{N_{\max}^i\}$ is independent, identically distributed, and integrable, then by the Strong Law of Large Numbers, $1/k \sum_{i=1}^k N_{\max}^i$ is bounded by an almost surely finite random variable. Therefore, the sequence $\{S_{n+1}/k\}$ is bounded by an almost surely finite random variable, say B_s , and from (36) we can write

$$\frac{\tilde{a}_k}{a_k} \geq \frac{A}{\bar{A}} (B_s)^{-\delta}.$$

Letting $A_l = A/(\bar{A}B_s^\delta)$, we have that $\tilde{a}_k \geq A_l a_k$, with $A_l > 0$ almost surely.

Using (b) of Lemma 3.8 and the above result we find

$$\sum_{k=1}^{\infty} \tilde{b}_k \geq \sum_{k=1}^{\infty} A_l a_k \geq A_l A \sum_{k=1}^{\infty} k^{-\delta} = \infty \quad \text{a.s.}$$

This verifies condition (A1).

Similarly, using (b) of Lemma 3.8 and the fact that $\tilde{a}_k \leq a_k$ we have

$$\sum_{k=1}^{\infty} \tilde{b}_k^2 \leq \sum_{k=1}^{\infty} B_N^2 \tilde{a}_k^2 \leq B_N^2 \sum_{k=1}^{\infty} a_k^2 \leq B_N^2 \bar{A}^2 \sum_{k=1}^{\infty} k^{-2\delta} < \infty.$$

This verifies condition (A2).

To show that (E1) is satisfied, we use (b) and (c) of Lemma 3.8, the fact that $\tilde{a}_k \leq a_k$, and (G3) to write

$$(37) \quad \begin{aligned} \sum_{k=1}^{\infty} \tilde{b}_k |E_{\tilde{\mathcal{F}}_k}(\tilde{\varepsilon}_{k+1})| &\leq B_N \sum_{k=1}^{\infty} \tilde{a}_k (B \tilde{a}_k + B_d \tilde{a}_k^d) \\ &\leq B_N \left(B \sum_{k=1}^{\infty} a_k^2 + B_d \sum_{k=1}^{\infty} a_k^{1+d} \right) \\ &\leq B_N B \bar{A}^2 \sum_{k=1}^{\infty} k^{-2\delta} + B_N B_d \sum_{k=1}^{\infty} a_k^{1+d}. \end{aligned}$$

The first term is finite since $\frac{1}{2} < \delta \leq 1$, and the second term is finite by (9). Thus, (37) is almost surely finite. This verifies condition (E1).

That condition (E2) is satisfied is a direct consequence of part (d) of Lemma 3.8.

All the conditions of Lemma 3.9 now hold, and the result follows. \square

We are now ready for step (M3), which provides the proof of Theorem 3.2.

Proof of Theorem 3.2. Fix $\epsilon > 0$. We wish to show that there exists an almost surely finite random variable M such that for all $n \geq M(\omega)$, $|\theta_n(\omega) - \theta^*| < \epsilon$.

For each $n \in \mathbb{N}$, let k_n be the index of the BP containing the n th customer. Let $F_n = S_{k_n} + 1$ and $G_n = S_{k_n+1} + 1$. So, F_n is the index of the first customer in the BP containing the n th customer, and G_n is the index of the first customer in the BP just after the one containing the n th customer. Clearly $G_n - F_n = \bar{N}_{k_n}$.

By Lemma 3.10, there exists an almost surely finite random variable M_1 such that if $k \geq M_1(\omega)$, then $|\theta_{S_k+1}(\omega) - \theta^*| < \epsilon/2$. Let $M_2 = S_{M_1} + 1$. Clearly, $M_2 < \infty$ almost surely. Then, for all $n \geq M_2(\omega)$, $k_n(\omega) \geq M_1(\omega)$ and so $|\theta_{F_n}(\omega) - \theta^*| < \epsilon/2$.

Now, by (13) in Lemma 3.7,

$$\begin{aligned} |\theta_n - \theta_{F_n}| &\leq \left| \sum_{i=F_n}^n \hat{a}_i^k (\hat{g}_{i+1} + C'(\theta_i)) \right| \\ &\leq \hat{a}_n^k \sum_{i=F_n}^{G_n-1} \frac{\hat{a}_i^k}{\hat{a}_n^k} (|\hat{g}_{i+1}| + |C'(\theta_i)|). \end{aligned}$$

For simplicity of notation, let $\bar{\phi}_j = \bar{\phi}_j^{k_n}$, $j \in \mathbb{N}$. Then, by properties (1) and (2) of $\{\hat{a}_n^k\}$, and equation (6),

$$\begin{aligned} |\theta_n - \theta_{F_n}| &\leq a_n \sum_{i=F_n}^{G_n-1} \frac{a_{F_n}}{a_{G_n-1}} \left(\sum_{j=1}^{i-F_n+1} \bar{\phi}_j + C'_M \right) \\ &\leq \bar{A} n^{-\delta} \sum_{i=F_n}^{G_n-1} \frac{\bar{A}}{\underline{A}} \left(\frac{F_n}{G_n-1} \right)^{-\delta} \sum_{j=1}^{\bar{N}_k} \bar{\phi}_j + C'_M \\ &= n^{-\delta} \frac{\bar{A}^2}{\underline{A}} \bar{N}_k \left(\sum_{j=1}^{\bar{N}_k} \bar{\phi}_j + C'_M \right) \left(\frac{G_n-1}{F_n} \right)^\delta. \end{aligned}$$

Now,

$$\frac{G_n-1}{F_n} = \frac{F_n + \bar{N}_k - 1}{F_n} \leq 1 + \frac{\bar{N}_k}{F_n} \leq 1 + \bar{N}_k \leq 2\bar{N}_k.$$

Thus,

$$\begin{aligned} |\theta_n - \theta_{F_n}| &\leq n^{-\delta} \frac{\bar{A}^2}{\underline{A}} \bar{N}_k \left(\sum_{j=1}^{\bar{N}_k} \bar{\phi}_j + C'_M \right) (2\bar{N}_k)^\delta \\ &\leq n^{-\delta} \frac{\bar{A}^2}{\underline{A}} \left(\sum_{j=1}^{N_{\max}^k} \bar{\phi}_j + C'_M \right) N_{\max}^k (2N_{\max}^k)^\delta. \end{aligned}$$

So, squaring and using Hölder's inequality and our known bounds, we obtain

$$\begin{aligned}
 E(\theta_n - \theta_{F_n})^2 &\leq n^{-2\delta} \frac{\bar{A}^4}{\underline{A}^2} E \left(\left(\sum_{j=1}^{N_{\max}^k} \bar{\phi}_j + C'_M \right)^2 (N_{\max}^k)^2 (2N_{\max}^k)^{2\delta} \right) \\
 (38) \quad &\leq n^{-2\delta} \frac{\bar{A}^4}{\underline{A}^2} \left(E \left(\sum_{j=1}^{N_{\max}^k} \bar{\phi}_j + C'_M \right)^6 \right)^{1/3} \left(E \left((N_{\max}^k)^3 (2N_{\max}^k)^{3\delta} \right) \right)^{2/3} \\
 &\leq n^{-2\delta} \frac{\bar{A}^4}{\underline{A}^2} (B'_\phi)^{1/3} (8B_N)^{2/3}.
 \end{aligned}$$

Now, by Markov's inequality,

$$(39) \quad P\{|\theta_n - \theta_{F_n}| \geq \epsilon/2\} \leq \frac{E(\theta_n - \theta_{F_n})^2}{(\epsilon/2)^2}.$$

Combining (38) with (39), we obtain

$$P\{|\theta_n - \theta_{F_n}| \geq \epsilon/2\} \leq n^{-2\delta} \hat{B}_\epsilon,$$

where

$$\hat{B}_\epsilon = \frac{\bar{A}^4}{\underline{A}^2} (B'_\phi)^{1/3} (8B_N)^{2/3} \frac{1}{(\epsilon/2)^2} < \infty.$$

Thus,

$$\sum_{n=1}^{\infty} P\{|\theta_n - \theta_{F_n}| \geq \epsilon/2\} \leq \hat{B}_\epsilon \sum_{n=1}^{\infty} n^{-2\delta} < \infty.$$

Hence, by the Borel–Cantelli lemma, $|\theta_n - \theta_{F_n}| < \epsilon/2$ except possibly for an almost surely finite (random) number of n , i.e., there exists an almost surely finite random variable M_3 such that for all $n \geq M_3(\omega)$, $|\theta_n(\omega) - \theta_{F_n}(\omega)| < \epsilon/2$.

Let $M = \max(M_2, M_3)$. Then, for all $n \geq M(\omega)$,

$$|\theta_n(\omega) - \theta^*| \leq |\theta_n(\omega) - \theta_{F_n}(\omega)| + |\theta_{F_n}(\omega) - \theta^*| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

which proves the result of Theorem 3.2. \square

Observe that step (M1) basically shows that the sequence $\{\tilde{\theta}_n\}$ behaves like an algorithm in which θ is updated after every BP, and step (M2) (via Lemma 3.9) is a convergence result for such an algorithm. This suggests that the rate of convergence of our algorithm is essentially that of one which updates after every BP. To be more specific, observe that from the proof of Lemma 3.8, the quantity \tilde{h}_{n+1} is “close” (differing only by $\tilde{\epsilon}_{n+1}$) to a quantity of the form

$$(40) \quad \frac{1}{E_{\tilde{\theta}_n}(N_n)} \sum_{i=1}^{N_n} \sum_{j=1}^i \phi_j^n(\tilde{\theta}_n) + C'(\tilde{\theta}_n).$$

The quantity in (40) is in fact an unbiased estimate of the derivative of J at $\tilde{\theta}_n$ (by Lemma 2.1), and can be used in an algorithm that updates θ after each BP (see

[8]). We may consider \tilde{h}_{n+1} to be an estimate of $dJ(\tilde{\theta}_n)/d\theta$, and the expression $\theta_{n+1} = \tilde{\pi}_{n+1}[\tilde{\theta}_n - \tilde{b}_n \tilde{h}_{n+1}]$ to be an algorithm which updates after every BP. If we replace \tilde{h}_{n+1} by the quantity in (40), then the resulting algorithm (which updates after every BP) should behave sample-pathwise similar to the subsequence $\{\tilde{\theta}_n\}$ of our original algorithm.

4. Extension to general update times. We now show how our algorithm can be extended so that parameter updates occur only at selected customers. For example, we may wish to update the parameter every time a fixed number of customers have been served, or, more generally, after the service of some random number of customers, determined by a prespecified criterion.

To model this situation let $H_\tau = \{\tau_1, \tau_2, \dots\}$ be a sequence of integer-valued random variables such that $\tau_1 < \tau_2 < \dots$ almost surely. The random variables τ_j will indicate for which customer the j th update will be performed. We think of H_τ as the (random) set of update indices, i.e., $n \in H_\tau$ means that we update before the service of the n th customer. For convenience we set $\tau_1 = 1$, so that θ_1 is set just before serving the first customer. Naturally, we require that $\{\tau_j = n\} \in \mathcal{F}_n$, i.e., that each τ_j is a stopping time with respect to $\{\mathcal{F}_n\}$.

For simplicity and convenience of analysis we will also make the following two assumptions:

(T1) There exists a positive integer B_τ such that for every j , $\tau_{j+1} - \tau_j \leq B_\tau$ almost surely; and

(T2) For almost all $\omega \in \Omega$, if $f_n(\omega) = 1$ (where f_n is defined by (1)), then there exists $j \in \mathbb{N}$ such that $\tau_j(\omega) = n$;

The first assumption simply ensures that updates are not too far apart. In practice this is a very mild restriction, since B_τ can be arbitrarily large. The second assumption ensures that a parameter update is always performed for the first customer in each BP. Since f_{n+1} is \mathcal{F}_{n+1} -measurable, i.e., whether the next customer to be served is the first in a BP is known, it is always possible to modify the sequence $\{\tau_j\}$, if necessary, so that this assumption is satisfied. At the cost of some further complication in the analysis the assumption can be weakened to the following condition:

(T2') There exists an integer t such that $f_{tn} = 1$ only if there exists $j \in \mathbb{N}$ such that $\tau_j = tn$ almost surely,

i.e., a parameter update is performed for the first customer in every t th BP. In this form, the assumption is very mild, since t can be arbitrarily large. However, for brevity we restrict our attention to assumption (T2) in our analysis.

Our modified algorithm can be summarized as follows. Between updates, the control parameter remains fixed, i.e., for each $j \in \mathbb{N}$, $\theta_{\tau_j} = \theta_{\tau_j+1} = \dots = \theta_{\tau_{j+1}-1}$. Let $\beta_j = \theta_{\tau_j}$ be the control parameter at the j th iteration of the algorithm. Then the sequence $\{\beta_j\}$ is generated by the recursion

$$(41) \quad \beta_{j+1} = \pi'_{j+1}[\beta_j - b_j \hat{f}_{j+1}],$$

where $\{b_j\}$ is the step-size sequence, $\pi'_{j+1}[\cdot]$ is a projection defined by

$$(42) \quad \pi'_{j+1}[x] = \begin{cases} \beta_j, & \text{if } x \notin D, \\ x, & \text{otherwise,} \end{cases}$$

and \hat{f}_{j+1} is an estimate of $dJ(\beta_j)/d\theta$ defined by

$$(43) \quad \hat{f}_{j+1} = \begin{cases} \sum_{n=\tau_j}^{\tau_{j+1}-1} (\hat{g}_{n+1} + C'(\theta_n)), & \text{if } \beta_{j-1} - b_{j-1} \hat{f}_j \in D \text{ or } f_{\tau_j} = 1; \\ \sum_{n=\tau_j}^{\tau_{j+1}-1} (\hat{g}_{n+1} + C'(\theta_n)) + \hat{f}_j, & \text{otherwise.} \end{cases}$$

At the j th iteration, our estimate of $dJ(\beta_j)/d\theta$ is simply the sum of the estimates obtained for each customer during the iteration. As before, in the event that a projection is required we add our derivative estimate to the next estimate, subject to resetting this accumulation after the start of the next BP.

The above algorithm encompasses a variety of interesting special cases. For example, by setting $\tau_j = j$ we obtain our original “one customer updating” algorithm. By setting $\tau_j = lj \wedge \min\{n : l(j-1) < n \leq lj \text{ \& } f_n = 1\}$, for some positive integer l , we obtain an algorithm that updates each time l customers are served, or a new BP is started. More generally, we may wish to update only after the derivative estimate meets some criterion, e.g., is deemed to be of sufficient accuracy or of the correct sign, subject to also satisfying (T1). As pointed out before, the original “one customer updating” algorithm can also be viewed as an algorithm that updates after a prespecified criterion is met, namely that $\theta_n - a_n \hat{h}_{n+1} \in D$ or that the next customer is the first in a BP.

Define the filtration $\{\mathcal{F}_j^\tau\}$ by $\mathcal{F}_j^\tau = \mathcal{F}_{\tau_j}$ (the natural filtration associated with the stopping times $\{\tau_j\}$). Let assumptions (G1')–(G4') be (G1)–(G4) with a_n , $\{\mathcal{F}_n\}$, \bar{A} and \underline{A} replaced by b_n , $\{\mathcal{F}_n^\tau\}$, \bar{B} and \underline{B} , respectively. Then, for the generalized algorithm above we have the following result.

THEOREM 4.1. *Suppose that $\{b_j\}$ satisfies (G1')–(G4'), and that (P1)–(P3), (S1)–(S4), and (C1)–(C3) are satisfied. Then, for the sequence $\{\beta_j\}$ defined by (41), (5), and (43), we have that $\beta_j \rightarrow \theta^*$ as $j \rightarrow \infty$ with probability one.*

Proof. To prove this result, we show that the θ_n are related via an algorithm that updates every customer, and that satisfies the conditions of Theorem 3.2. To this end, we first define the projection $\pi_{n+1}[\cdot]$ as follows:

$$(44) \quad \pi_{n+1}[x] = \begin{cases} \theta_n, & \text{if } x \notin D \text{ or } n \notin H_\tau, \\ x, & \text{otherwise.} \end{cases}$$

Correspondingly, we define the projection indicator sequence $\{p_n\}$ by

$$(45) \quad p_{n+1} = \begin{cases} 1, & \text{if } n+1 \notin H_\tau \text{ or } \theta_n - a_n \hat{h}_{n+1} \notin D, \\ 0, & \text{otherwise.} \end{cases}$$

Then define the sequence $\{a_n\}$ by $a_{\tau_j} = a_{\tau_j+1} = \cdots = a_{\tau_{j+1}-1} = b_j$, $j \in \mathbb{N}$. It is now easy to see that the θ_n satisfy the recursion

$$\theta_{n+1} = \pi_{n+1}[\theta_n - a_n \hat{h}_{n+1}],$$

where $\{\hat{h}_{n+1}\}$ is defined by (7) with the sequence $\{p_n\}$ defined above.

Now, the proof of Theorem 3.2 does not depend on the sequence $\{p_n\}$, except through Lemma 3.7. Note, however, that Lemma 3.7 is a purely algebraic result and holds for an arbitrary sequence $\{p_n\}$. Therefore, provided the assumptions of Theorem 3.2 hold, our result follows. Since assumptions (P1)–(P3), (S1)–(S4), and (C1)–(C3) are already assumed to hold, then it remains only to show that $\{a_n\}$ satisfies (G1)–(G4).

Assumptions (G1) and (G2) clearly hold for $\{a_n\}$. To verify (G3), for each n define $\sigma_n \triangleq \max\{j : n \geq \tau_j\}$, so that $a_n = b_{\sigma_n}$. By assumption, $j \leq \tau_j \leq jB_\tau$, and therefore $\sigma'_n \leq \sigma_n \leq n$, where $\sigma'_n = \max(1, (n/B_\tau) - 1)$. By (G2'), we may write $b_n \leq a_n \leq b_{\sigma'_n}$. By (G3'), we have

$$\underline{B}n^{-\delta} \leq a_n \leq \bar{B}(\sigma'_n)^{-\delta}.$$

Since $\sigma'_n \geq n/(2B_\tau)$, then

$$\underline{B}n^{-\delta} \leq a_n \leq \bar{B} \left(\frac{n}{2B_\tau} \right)^{-\delta}.$$

Set $\underline{A} = \underline{B}$ and $\bar{A} = \bar{B}(2B_\tau)^\delta$. Then (G3) is satisfied. To see that (G4) holds, we note that for $n+1 \in H_\tau$, $a_n = b_{\sigma_n}$ and $a_{n+1} = b_{\sigma_{n+1}}$, so that by (G4'),

$$\frac{1}{a_{n+1}} - \frac{1}{a_n} \leq B_a.$$

For $n+1 \notin H_\tau$, $a_{n+1} = a_n$, and so the above inequality remains true. Hence (G4) holds.

Therefore, by Theorem 3.2, $\theta_n \rightarrow \theta^*$ almost surely. Since $\{\beta_j\}$ is a subsequence of $\{a_n\}$, we also have $\beta_j \rightarrow \theta^*$ almost surely, which completes the proof. \square

5. More general performance measures. The proof of our main result can be extended to more general performance measures. For example, consider performance measures of the type $J(\theta) = E(F(t(\theta), \theta))$, where $F: \mathbb{R}_+ \times D \rightarrow \mathbb{R}$ is a given “nice” function and $t(\theta)$ is a random variable that has the distribution of the steady-state system time. By a “nice” function we mean that F has partial derivatives $\partial_t F$ and $\partial_\theta F$ that are bounded and Lipschitz continuous on $\mathbb{R}_+ \times D$. In Appendix A we have shown that

$$\frac{d}{d\theta} J(\theta) = \frac{1}{E(N_1(\theta))} E \left(\sum_{i=1}^{N_1(\theta)} \partial_t F(t_i(\theta), \theta) \sum_{j=1}^i \psi_\theta(X_j(\theta)) + \partial_\theta F(t_i(\theta), \theta) \right),$$

where $\{t_n(\theta)\}$ is the sequence of system times of customers, defined recursively by

$$t_n = \begin{cases} X_n(\theta) - A_{n-1}, & \text{if } f_n = 1, \\ t_{n-1} + X_n(\theta) - A_{n-1}, & \text{otherwise,} \end{cases}$$

with $A_0 = 0$. Correspondingly, we define the sequence of derivative estimates $\{\hat{h}_n\}$ by

$$(46) \quad \hat{h}_{n+1} = \begin{cases} \partial_t F(\hat{t}_{i+1}, \theta_n) \hat{g}_{n+1} + \partial_\theta F(\hat{t}_{i+1}, \theta_n), & \text{if } p_n = 0 \text{ or } f_n = 1, \\ \partial_t F(\hat{t}_{i+1}, \theta_n) \hat{g}_{n+1} + \partial_\theta F(\hat{t}_{i+1}, \theta_n) + \hat{h}_n, & \text{otherwise,} \end{cases}$$

where $\{\hat{t}_n\}$ is defined recursively by

$$(47) \quad \hat{t}_{n+1} = \begin{cases} X_n(\theta_n) - A_{n-1}, & \text{if } f_n = 1, \\ \hat{t}_n + X_n(\theta_n) - A_{n-1}, & \text{otherwise.} \end{cases}$$

Under suitable assumptions, the approach of §3.5 can be used to show that the algorithm defined by (2), (5), (8), (46), and (47) converges almost surely. The proof involves replacing all occurrences of terms such as $\hat{g}_{n+1} + C'(\theta_n)$ by $\partial_t F(\hat{t}_{i+1}, \theta_n) \hat{g}_{n+1} + \partial_\theta F(\hat{t}_{i+1}, \theta_n)$ in the proofs of Lemma 3.8 and Theorem 3.2, and reestablishing the bounds obtained in these proofs.

6. Conclusions. We have proved convergence of a stochastic optimization algorithm using IPA for a GI/G/1 queue with general update times. Simulation plots of our algorithm have been presented in [15]. Our proof made use of certain assumptions on the GI/G/1 queue—roughly, that the service-time distribution has finite moments up to a sufficiently high order, and that the interarrival-time distribution has a bounded hazard rate. We view the latter as the more restrictive of the two main assumptions. This assumption was used only to prove Lemma 3.6. It could be relaxed if an alternative proof for Lemma 3.6 can be found.

Our approach may provide a fruitful line of investigation into the convergence of algorithms applied to networks of queues. Of course, there are many more technical complexities in this setting. This is a difficult but important practical problem.

Appendix A. Proof of Lemma 2.1. We will prove a general result, of which Lemma 2.1 is a special case. For this, let $F : \mathbb{R}_+ \times D \rightarrow \mathbb{R}$ be a function with partial derivatives $\partial_t F$ and $\partial_\theta F$ that are bounded on $\mathbb{R}_+ \times D \rightarrow \mathbb{R}$, and let $t(\theta)$ be a random variable that has the distribution of the steady-state system time. We will show that under the assumptions of Lemma 2.1,

(48)

$$\frac{d}{d\theta} E(F(t(\theta), \theta)) = \frac{1}{E(N_1(\theta))} E \left(\sum_{i=1}^{N_1(\theta)} \partial_t F(t_i(\theta), \theta) \sum_{j=1}^i \psi_\theta(X_j(\theta)) + \partial_\theta F(t_i(\theta), \theta) \right).$$

If we let F be the function $(t, \theta) \mapsto t$, then $\partial_t F = 1$, $\partial_\theta F = 0$, and (48) reduces to the case of Lemma 2.1.

For each $\theta \in D$, define a sequence $\{R_k(\theta)\}$ by $R_k(\theta) = \sum_{i=1}^k N_i(\theta)$, $k \geq 1$, and $R_0(\theta) = 0$. So $R_{k-1} + 1$ is simply the index of the first customer in the k th BP. Define a filtration $\{\mathcal{G}_n\}$ by $\mathcal{G}_1 = \sigma(A_1, \mathbf{X}_1)$, and for $n > 1$, $\mathcal{G}_n = \sigma(\mathcal{G}_{n-1}, A_n, \mathbf{X}_n)$. It can be shown that for each θ and k , $R_k(\theta)$ is a stopping time with respect to $\{\mathcal{G}_n\}$.

Note that for each $k \in \mathbb{N}$, $R_k(\theta)$ is simply the smallest positive integer satisfying

$$(49) \quad \sum_{i=R_{k-1}(\theta)+1}^{R_k(\theta)} X_i(\theta) < \sum_{i=R_{k-1}(\theta)+1}^{R_k(\theta)} A_i.$$

Let $N_{\max}^k = N_k(\theta_{\max})$. For each n , let $t_n(\theta)$ be the system time of the n th customer (with control parameter value θ). Specifically, if the n th customer is within the k th BP, then $t_n(\theta)$ is given by

$$t_n(\theta) = \sum_{i=R_{k-1}(\theta)+1}^n X_i(\theta) - \sum_{i=R_{k-1}(\theta)+1}^{n-1} A_i.$$

It is easy to see that the stopping time $R_k(\theta)$ are regeneration points for the process $\{t_n(\theta)\}$ (see [16]). Now, from the previous argument, the $R_k(\theta_{\max})$ are stopping times with respect to $\{\mathcal{G}_n\}$. We claim that the $R_k(\theta_{\max})$ are also regeneration points for the process $\{t_n(\theta)\}$. To see this, it suffices to show that for each k , there exists a random variable K_k such that $R_k(\theta_{\max}) = R_{K_k}(\theta)$, i.e., $\{R_k(\theta_{\max})\}$ is a subsequence of $\{R_k(\theta)\}$. It will suffice for us to show that there exists K_1 such that $R_1(\theta_{\max}) = R_{K_1}(\theta)$ (note that $R_1(\theta_{\max}) = N_{\max}^1$). To this end, let $K_1 = 1 + \max\{j \in \mathbb{N} : R_j(\theta) <$

$N_{\max}^1\}$. Now, $R_{K_1-1}(\theta) < N_{\max}^1$, and therefore

$$(50) \quad \sum_{i=1}^{R_{K_1-1}(\theta)} X_i(\theta_{\max}) \geq \sum_{i=1}^{R_{K_1-1}(\theta)} A_i.$$

Also, we note that N_{\max}^1 is the smallest positive integer satisfying

$$(51) \quad \sum_{i=1}^{N_{\max}^1} X_i(\theta_{\max}) < \sum_{i=1}^{N_{\max}^1} A_i.$$

Subtracting (50) from (51), we get that

$$\sum_{i=R_{K_1-1}(\theta)+1}^{N_{\max}^1} X_i(\theta_{\max}) < \sum_{i=R_{K_1-1}(\theta)+1}^{N_{\max}^1} A_i.$$

Since for all i , $X_i(\theta) \leq X_i(\theta_{\max})$, then we also have that

$$\sum_{i=R_{K_1-1}(\theta)+1}^{N_{\max}^1} X_i(\theta) < \sum_{i=R_{K_1-1}(\theta)+1}^{N_{\max}^1} A_i.$$

Since $R_{K_1}(\theta)$ is the smallest positive integer satisfying (49), then $R_k(\theta) \leq N_{\max}^1$. But by definition of K_1 , $R_{K_1}(\theta) \geq N_{\max}^1$. Hence, $R_{K_1}(\theta) = N_{\max}^1$.

Now, since the stopping times $R_k(\theta_{\max})$ are regeneration points for the process $\{t_n(\theta)\}$ (and hence also $\{F(t_n(\theta), \theta)\}$), then from the theory of regenerative systems (e.g., see [16]),

$$(52) \quad E(F(t(\theta), \theta)) = \frac{1}{E(N_{\max}^1)} E \left(\sum_{i=1}^{N_{\max}^1} F(t_i(\theta), \theta) \right).$$

Since each X_n is differentiable with respect to θ on D , then we claim that $F(t_n, \cdot)$ is also differentiable with respect to θ on D . It suffices to show that t_n is differentiable on D . To see this, first fix $\omega \in \Omega$. For each k , there exists an open interval $I_k(\omega)$ containing θ such that $N_k(\omega)$ is constant over $I_k(\omega)$. Therefore, there exists an open interval $I(\omega)$ containing θ such that for all $k \leq K_1(\omega)$, $R_k(\omega)$ is constant over $I(\omega)$. Consider the n th customer, where $n \leq N_{\max}^1(\omega)$. Suppose it is within the $k_n(\omega)$ th BP. Then clearly $k_n(\omega) \leq N_{\max}^1(\omega)$. Since (henceforth dropping the argument ω)

$$t_n(\theta) = \sum_{i=R_{k_n-1}+1}^n X_i(\theta) - \sum_{i=R_{k_n-1}+1}^{n-1} A_i,$$

then

$$\begin{aligned} \frac{dt_n}{d\theta}(\theta) = \lim_{\delta \rightarrow 0} \frac{1}{\delta} & \left(\sum_{i=R_{k_n-1}(\theta+\delta)+1}^n X_i(\theta+\delta) - \sum_{i=R_{k_n-1}(\theta)+1}^{n-1} X_i(\theta) \right. \\ & \left. - \sum_{i=R_{k_n-1}(\theta+\delta)+1}^{n-1} A_i + \sum_{i=R_{k_n-1}(\theta)+1}^{n-1} A_i \right). \end{aligned}$$

For all δ small enough such that $\theta + \delta \in I$, $R_{k_n-1}(\theta + \delta) = R_{k_n-1}(\theta)$ and hence

$$\sum_{i=R_{k_n-1}(\theta+\delta)+1}^n X_i(\theta + \delta) - \sum_{i=R_{k_n-1}(\theta)+1}^n X_i(\theta) = \sum_{i=R_{k_n-1}(\theta)+1}^n X_i(\theta + \delta) - X_i(\theta),$$

and, similarly,

$$\sum_{i=R_{k_n-1}(\theta+\delta)+1}^{n-1} A_i - \sum_{i=R_{k_n-1}(\theta)+1}^{n-1} A_i = 0.$$

Therefore,

$$\begin{aligned} \frac{dt_n}{d\theta}(\theta) &= \lim_{\delta \rightarrow 0} \sum_{i=R_{k_n-1}(\theta)+1}^n \frac{X_i(\theta + \delta) - X_i(\theta)}{\delta} \\ &= \sum_{i=R_{k_n-1}(\theta)+1}^n \lim_{\delta \rightarrow 0} \frac{X_i(\theta + \delta) - X_i(\theta)}{\delta} \\ &= \sum_{i=R_{k_n-1}+1}^n \frac{dX_i}{d\theta}(\theta). \end{aligned}$$

Since by assumption (P4), $dX_i(\theta)/d\theta = \psi_\theta(X_i(\theta))$, then we may write

$$\frac{dt_n}{d\theta}(\theta) = \sum_{i=R_{k_n-1}+1}^n \psi_\theta(X_i(\theta)).$$

Fix $\theta \in D$. We now claim that

$$\frac{d}{d\theta} E \left(\sum_{i=1}^{N_{\max}^1} F(t_i(\theta), \theta) \right) = E \left(\sum_{i=1}^{N_{\max}^1} \frac{d}{d\theta} F(t_i(\theta), \theta) \right).$$

To show this, we first define, for each $j \in \mathbb{N}$, $\bar{\phi}_j \triangleq \sup_{\theta \in D} (dX_j/d\theta)(\theta)$. Since $\partial_t F$ and $\partial_\theta F$ are both bounded, then there exists a finite constant B_F such that $\partial_t F, \partial_\theta F \leq B_F$. Let

$$Y = \sum_{i=1}^{N_{\max}^1} B_F \left(1 + \sum_{j=1}^i \bar{\phi}_j \right).$$

It is clear, by assumption (P3), that for each $i \leq N_{\max}$,

$$\left| \frac{dt_i}{d\theta}(\theta) \right| \leq \sum_{j=1}^i |\psi_\theta(X_j(\theta))|.$$

Therefore, we have that for all $\theta \in D$,

$$\begin{aligned} \left| \sum_{i=1}^{N_{\max}^1} \frac{d}{d\theta} F(t_i(\theta), \theta) \right| &= \left| \sum_{i=1}^{N_{\max}^1} \partial_t F(t_i(\theta), \theta) \frac{dt_i}{d\theta}(\theta) + \partial_\theta F(t_i(\theta), \theta) \right| \\ &\leq \sum_{i=1}^{N_{\max}^1} \left(B_F \sum_{j=1}^i |\psi_\theta(X_j(\theta))| + B_F \right) \leq \sum_{i=1}^{N_{\max}^1} B_F \left(1 + \sum_{j=1}^i \bar{\phi}_j \right) = Y. \end{aligned}$$

Now, we can write

$$\begin{aligned} |Y| &\leq \sum_{i=1}^{N_{\max}^1} B_F \left(1 + \sum_{j=1}^{N_{\max}^1} \bar{\phi}_j \right) \\ &= B_F N_{\max}^1 + B_F N_{\max}^1 \sum_{j=1}^{N_{\max}^1} (\bar{\phi}_j - E(\bar{\phi}_1)) + B_F (N_{\max}^1)^2 E(\bar{\phi}_1). \end{aligned}$$

By one of Wald's identities (see [17, p. 460]) and assumption (P4)

$$E \left(\sum_{j=1}^{N_{\max}^1} \bar{\phi}_j - E(\bar{\phi}_1) \right)^2 = E(N_{\max}^1) \text{Var}(\bar{\phi}_1) < \infty.$$

From assumption (P2) and Lemma 3.3, we have that

$$\begin{aligned} E(|Y|) &\leq B_F E(N_{\max}^1) + B_F \sqrt{E(N_{\max}^1)^2 E \left(\sum_{j=1}^{N_{\max}^1} \bar{\phi}_j - E(\bar{\phi}_1) \right)^2} \\ &\quad + B_F E(N_{\max}^1)^2 E(\bar{\phi}_1) \\ &< \infty, \end{aligned}$$

which shows that Y is integrable. Therefore, by continuity of $\sum_{i=1}^{N_{\max}^1} F(t_i(\cdot), \cdot)$ and [18, p. 46], we have that

$$(53) \quad \frac{d}{d\theta} E \left(\sum_{i=1}^{N_{\max}^1} F(t_i(\theta), \theta) \right) = E \left(\sum_{i=1}^{N_{\max}^1} \frac{d}{d\theta} F(t_i(\theta), \theta) \right).$$

Therefore, combining (53) with (52) we have that

$$(54) \quad \frac{d}{d\theta} F(t(\theta), \theta) = \frac{1}{E(N_{\max}^1)} E \left(\sum_{i=1}^{N_{\max}^1} \frac{d}{d\theta} F(t_i(\theta), \theta) \right).$$

Recall the definition for the sequence $\{R_k(\theta)\}$, where θ is the value previously fixed (we will hereafter drop the argument θ and simply write $\{R_k\}$). Since $\{R_k\}$ is a sequence of stopping times relative to $\{\mathcal{G}_n\}$, we may define a sequence of σ -algebras $\{\mathcal{H}_k\}$ by $\mathcal{H}_k = \mathcal{G}_{R_k}$, where $\mathcal{G}_{R_k} \triangleq \{A \in \mathcal{G} : A \cap \{R_k = i\} \in \mathcal{G}_i \text{ for all } i \in \mathbb{N}\}$ (see, e.g., [17, p. 449]). As $R_k \leq R_{k+1}$ almost surely for all $k \in \mathbb{N}$, then $\{\mathcal{H}_k\}$ is a filtration (see, e.g., [17, p. 455]), and $\{R_k\}$ is adapted to $\{\mathcal{H}_k\}$. It can be shown that K_1 is a stopping time with respect to $\{\mathcal{H}_k\}$.

Consider (54). Now, we may write

$$N_{\max}^1 = R_{K_1} = \sum_{k=1}^{K_1} N_k.$$

Therefore,

$$E(N_{\max}^1) = E \left(\sum_{k=1}^{K_1} N_k \right).$$

Since $\{N_k\}$ is an independent and identically distributed sequence adapted to $\{\mathcal{H}_k\}$, and K_1 is a stopping time with respect to $\{\mathcal{H}_k\}$, then by one of Wald's identities (see [17, p. 460]), we may write

$$(55) \quad E \left(\sum_{k=1}^{K_1} N_k \right) = E(K_1)E(N_1).$$

Similarly, we may write

$$\sum_{i=1}^{N_{\max}^1} \frac{d}{d\theta} F(t_i(\theta), \theta) = \sum_{i=1}^{R_{K_1}} \frac{d}{d\theta} F(t_i(\theta), \theta) = \sum_{k=1}^{K_1} \sum_{i=R_{k-1}+1}^{R_k} \frac{d}{d\theta} F(t_i(\theta), \theta).$$

We now note that the sequence of random variables

$$\left\{ \sum_{i=R_{k-1}+1}^{R_k} \frac{d}{d\theta} F(t_i(\theta), \theta) \right\}_{k \in \mathbb{N}}$$

is an independent and identically distributed sequence adapted to $\{\mathcal{H}_k\}$, and again by one of Wald's identities, we may write

$$(56) \quad E \left(\sum_{i=1}^{N_{\max}^1} \frac{d}{d\theta} F(t_i(\theta), \theta) \right) = E(K_1)E \left(\sum_{i=1}^{N_1} \frac{d}{d\theta} F(t_i(\theta), \theta) \right).$$

Combining (55) and (56) with (54) we obtain

$$\begin{aligned} \frac{d}{d\theta} F(t(\theta), \theta) &= \frac{1}{E(N_1)} E \left(\sum_{i=1}^{N_1} \frac{d}{d\theta} F(t_i(\theta), \theta) \right) \\ &= \frac{1}{E(N_1)} E \left(\sum_{i=1}^{N_1} \partial_t F(t_i(\theta), \theta) \sum_{j=1}^i \psi_{\theta}(X_j(\theta)) + \partial_{\theta} F(t_i(\theta), \theta) \right) \end{aligned}$$

which completes the proof. \square

Appendix B. Proof of Proposition 3.1. For simplicity of notation, let $X(\theta) = X_m^k(\theta)$, $K = K_m^k$, $\bar{K} = \bar{K}_m^k$, $\phi = \phi_m^k$, and $\bar{\phi} = \bar{\phi}_m^k$. Let $\theta_1, \theta_2 \in D$. Then

$$\begin{aligned} |\phi(\theta_2) - \phi(\theta_1)| &= |\psi_{\theta_2}(X(\theta_2)) - \psi_{\theta_1}(X(\theta_1))| \\ &= |\psi_{\theta_2}(X(\theta_2)) - \psi_{\theta_1}(X(\theta_2)) + \psi_{\theta_1}(X(\theta_2)) - \psi_{\theta_1}(X(\theta_1))| \\ &\leq K_1(X(\theta_2))|\theta_2 - \theta_1| + K_2(\theta_1)|X(\theta_2) - X(\theta_1)|. \end{aligned}$$

By the Mean Value Theorem,

$$\begin{aligned} |X(\theta_2) - X(\theta_1)| &= \left| \frac{\partial X}{\partial \theta}(\tilde{\theta}) \right| |\theta_2 - \theta_1| \\ &= |\psi_{\tilde{\theta}}(X(\tilde{\theta}))| |\theta_2 - \theta_1| \\ &= |\phi(\tilde{\theta})| |\theta_2 - \theta_1| \\ &\leq \bar{\phi} |\theta_2 - \theta_1|, \end{aligned}$$

where $\tilde{\theta} = \theta_1 + \xi(\theta_2 - \theta_1)$ with $0 \leq \xi \leq 1$. So,

$$\begin{aligned} |\phi(\theta_2) - \phi(\theta_1)| &\leq K_1(X(\theta_2))|\theta_2 - \theta_1| + \bar{L}\bar{\phi}|\theta_2 - \theta_1| \\ &\leq (\bar{K} + \bar{L}\bar{\phi})|\theta_2 - \theta_1| \\ &= K|\theta_2 - \theta_1|, \end{aligned}$$

where $K \triangleq \bar{K} + \bar{L}\bar{\phi}$. Now,

$$\begin{aligned} E(K^2) &= E(\bar{K}^2) + 2\bar{L}E(\bar{K}\bar{\phi}) + \bar{L}^2E(\bar{\phi}^2) \\ &\leq E(\bar{K}^2) + 2\bar{L}\sqrt{E(\bar{K}^2)}\sqrt{E(\bar{\phi}^2)} + \bar{L}^2E(\bar{\phi}^2) < \infty \end{aligned}$$

which completes the proof. \square

Appendix C. Proof of Lemma 3.4. It suffices to prove the result for $r = p$. Fix $k \in \mathbb{N}$ and $\theta \in D$. Define a filtration $\{\mathcal{F}_m^k\}$ by $\mathcal{F}_m^k = \sigma(X_i^k, A_i^k, i = 1, \dots, m)$. Then, $N_k(\theta)$ is a stopping time with respect to $\{\mathcal{F}_m^k\}$, since

$$\{N_k(\theta) = m\} = \left(\bigcap_{j=1}^{m-1} \left\{ \sum_{i=1}^{j-1} (X_i^k(\theta) - A_i^k) \geq 0 \right\} \right) \cap \left\{ \sum_{i=1}^m (X_i^k(\theta) - A_i^k) < 0 \right\}$$

and the sets on the right-hand side are in \mathcal{F}_m^k . Now, since $\bar{\phi}_m^k$ is a function of X_m^k , then it is clearly \mathcal{F}_m^k -measurable and independent of \mathcal{F}_{m-1}^k . Therefore, by [14, Thm. 5.2, p. 22],

$$E \left(\sum_{i=1}^{N_k(\theta)} \bar{\phi}_i^k \right)^r \leq B_r' E(\bar{\phi}_1^1)^r E(N_k(\theta))^r,$$

where B_r' is a numerical constant depending only on r . Therefore, if we let $B_\phi = B_r' E(\bar{\phi}_1^1)^r B_N$ then by Lemma 3.3, we may write

$$E \left(\sum_{i=1}^{N_k(\theta)} \bar{\phi}_i^k \right)^r \leq B_\phi.$$

This completes the proof. \square

Appendix D. Proof of Lemma 3.5. Fix k and $\theta \in D$. Since $E(N_1(\theta)) < \infty$ and $E(K_1^1)^2 < \infty$, then by one of Wald's identities (see [17, p. 460]),

$$(57) \quad E \left(\sum_{i=1}^{N_k(\theta)} K_i^k - E(K_i^k) \right)^2 = E(K_1^1 - E(K_1^1))^2 E(N_1(\theta)).$$

Now,

$$\begin{aligned} &E \left(\sum_{i=1}^{N_k(\theta)} K_i^k \right)^2 \\ &= E \left(\sum_{i=1}^{N_k(\theta)} K_i^k - E(K_i^k) \right)^2 + 2E \left(N_k(\theta) \sum_{i=1}^{N_k(\theta)} K_i^k - E(K_i^k) \right) E(K_1^1) \end{aligned}$$

$$\begin{aligned}
& + (E(K_1^1))^2 E(N_1(\theta))^2 \\
& \leq E \left(\sum_{i=1}^{N_k(\theta)} K_i^k - E(K_i^k) \right)^2 + 2E(K_1^1) \sqrt{E(N_k(\theta))^2} \sqrt{E \left(\sum_{i=1}^{N_k(\theta)} K_i^k - E(K_i^k) \right)^2} \\
& \quad + (E(K_1^1))^2 E(N_1(\theta))^2.
\end{aligned}$$

Therefore, by (57) and using Lemma 3.3,

$$\begin{aligned}
& E \left(\sum_{i=1}^{N_k(\theta)} K_i^k \right)^2 \\
& \leq E(K_1^1 - E(K_1^1))^2 E(N_1(\theta)) + 2E(K_1^1) \sqrt{B_N} \sqrt{E(K_1^1 - E(K_1^1))^2 E(N_1(\theta))} \\
& \quad + (E(K_1^1))^2 B_N \\
& \leq B_K,
\end{aligned}$$

where

$$B_K = E(K_1^1 - E(K_1^1))^2 B_N + 2E(K_1^1) \sqrt{B_N} \sqrt{E(K_1^1 - E(K_1^1))^2 B_N} + (E(K_1^1))^2 B_N < \infty,$$

which concludes the proof. \square

REFERENCES

- [1] P. W. GLYNN, *Optimization of stochastic systems*, in Proc. 1986 Winter Simulation Conf., Washington, DC, pp. 52–59, 1986.
- [2] M. S. MEKTON, *Optimization in simulation: A survey of recent results*, in Proc. 1987 Winter Simulation Conf., Atlanta, GA, pp. 58–67, 1987.
- [3] M. C. FU AND Y.-C. HO, *Using perturbation analysis for gradient estimation, averaging and updating in a stochastic approximation algorithm*, in Proc. 1988 Winter Simulation Conf., San Diego, CA, pp. 509–517, 1988.
- [4] R. SURI AND M. A. ZAZANIS, *Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue*, Management Sci., 34 (1988), pp. 39–64.
- [5] Y. WARDI, *Simulation-Based Stochastic Algorithms for Optimizing GI/G/1 Queues*, preprint, Dept. of Industrial Engr., Ben Gurion University of the Negev, 1988.
- [6] R. SURI AND Y. T. LEUNG, *Single run optimization of discrete event simulations—An empirical study using the M/M/1 queue*, IIE Transactions, 21 (1989), pp. 35–49.
- [7] M. C. FU, *Convergence of a stochastic approximation algorithm for the GI/G/1 queue using infinitesimal perturbation analysis*, J. Optim. Theory Appl., 65 (1990), pp. 149–160.
- [8] E. K. P. CHONG AND P. J. RAMADGE, *Convergence of recursive optimization algorithms using infinitesimal perturbation analysis estimates*, Discrete Event Dynamic Systems: Theory and Applications, 1 (1992), pp. 339–372.
- [9] F. J. VÁZQUEZ-ABAD, *Stochastic Recursive Algorithms for Optimal Routing in Queueing Networks*, Ph.D. Thesis, Division of Applied Mathematics, Brown University, May 1989.
- [10] P. L'ECUYER, N. GIROUX, AND P. GLYNN, *Stochastic Optimization by Simulation: Convergence Proofs and Experimental Results for the GI/G/1 Queue*, manuscript, Département d'I.R.O., Université de Montréal, Nov. 1990.
- [11] M. METIVIER AND P. PRIOURET, *Applications of a Kushner and Clark lemma to general classes of stochastic algorithms*, IEEE Trans. Inform. Theory, IT-30 (1984), pp. 140–151.
- [12] P. KONSTANTOPOULOS AND M. ZAZANIS, *Sensitivity analysis for stationary and ergodic queue*, INRIA-Sophia Antipolis, Rapports de Recherche No. 1315, Programme 3, Réseaux et Systèmes Répartis, 1990.
- [13] H. THORISSON, *The queue GI/G/1: Finite moments of the cycle variables and uniform rates of convergence*, Stochastic Process. Appl., 19 (1985), pp. 85–99.

- [14] A. GUT, *Stopped Random Walks: Limit Theorems and Applications*, Springer-Verlag, New York, Berlin, 1988.
- [15] E. K. P. CHONG AND P. J. RAMADGE, *On a stochastic optimization algorithm using IPA which updates after every customer*, in Proc. 28th Allerton Conf. on Communication, Control, and Computing, Monticello, IL, Oct. 1990, pp. 658–667.
- [16] G. S. SHEDLER, *Regeneration and Networks of Queues*, Springer-Verlag, New York, Berlin, 1987.
- [17] A. N. SHIRYAYEV, *Probability*, Springer-Verlag, New York, Berlin, 1984.
- [18] R. G. BARTLE, *The Elements of Integration*, John Wiley, New York, 1966.

A SET INDUCED NORM APPROACH TO THE ROBUST CONTROL OF CONSTRAINED SYSTEMS*

MARIO SZNAIER[†]

Abstract. Most realistic control problems involve both some type of time-domain constraints and model uncertainty. However, the majority of controller design procedures currently available focus only on one aspect of the problem, with only a handful of methods capable of simultaneously addressing both issues. Recently, it has been proposed to address this class of problems by using a “constrained robustness measure,” generated by a constraint-set induced operator norm, to assess the stability properties of a family of systems. In this paper the properties of this constrained-robustness measure are explored and the theoretical framework is extended to include control as well as state constraints. These results are applied to the problem of designing fixed-order stabilizing feedback controllers for systems subject to structured parametric model uncertainty and time-domain constraints.

Key words. constrained systems, discrete time systems, robust design techniques, robust stability

AMS(MOS) subject classifications. 93, 93B35, 93B51, 93C55, 93D09

1. Introduction. A substantial number of control problems can be summarized as the problem of designing a controller capable of achieving acceptable performance under system uncertainty and design constraints. This statement looks deceptively simple, but even in the case where the system under consideration is linear, the problem is far from solved. Several methods have been proposed recently to deal with constrained control problems under the assumption of exact knowledge of the model (see [1] and references therein). However, such an assumption can be too restrictive, ruling out cases where good qualitative models of the plant are available but the numerical values of various parameters are unknown or even change during operation.

On the other hand, during the last decade a considerable amount of time has been spent analyzing the question of whether some relevant properties of a system (most notably asymptotic stability) are preserved under the presence of unknown perturbations. This research effort has led to procedures for designing “robust” controllers, capable of achieving desirable properties under various classes of plant perturbations while, at the same time, satisfying frequency-domain constraints. However, most of these design procedures cannot accommodate directly time-domain constraints (which precludes their use in cases where there exist physically motivated “hard” bounds on the states or control effort), although some progress has been recently made in this direction [2]–[5].

In [6], [7] we proposed to approach time-domain constrained systems using an operator norm-theoretic approach. We introduced a simple robustness measure that indicated how well the family of systems under consideration satisfied a given set of time-domain constraints and we proposed a design method yielding controllers that maximized this robustness measure. In this paper we extend our formalism to include

*Received by the editors August 7, 1991; accepted for publication (in revised form) March 10, 1992. This work was supported in part by a grant from the Division of Sponsored Research, University of Central Florida.

[†]Department of Electrical Engineering, University of Central Florida, Orlando, Florida 32816-0450.

control as well as a more general description of state constraints, and we explore the properties of the resulting constrained robustness measure. These theoretical results are applied to the problem of designing stabilizing controllers for systems subject to structured parametric model uncertainty and time-domain constraints. We show that in cases of practical interest the synthesis problem can be reduced to a convex, albeit in general nondifferentiable, optimization problem. We believe that the results presented here will provide a useful new approach for addressing more realistic control design problems.

The paper is organized as follows: In §2 we introduce the concepts of *constrained stability* and *robust constrained stability* and we use these concepts to give a formal definition of the *robust constrained stability analysis* and *robust constrained stability design* problems. The *analysis* problem is studied in §3 where we give necessary and sufficient conditions for constrained stability. We use these results to define a constrained robustness measure and we show that, under mild assumptions, this measure is a continuous, concave function of the dynamics of the system. In §4 we apply the results of §3 to the *design* problem and we show that in cases of practical interest our approach yields a well-behaved optimization problem. Finally, in §5 we summarize our results and indicate directions for future research.

2. Definitions and background results. In this section we give a formal definition of the robust constrained control problem. We begin by introducing several required concepts and preliminary results.

2.1. Preliminary definitions.

DEFINITION 1. Consider the linear, time invariant, discrete time, autonomous system modeled by the difference equation

$$(S^a) \quad \underline{x}_{k+1} = A\underline{x}_k, \quad k = 0, 1, \dots$$

subject to the constraint

$$(1) \quad \underline{x} \in \mathcal{G} \subset R^n$$

where $A \in R^{n \times n}$ and where \underline{x} indicates x is a vector quantity. The system (S^a) is *constrained stable* if for any point $\underline{\tilde{x}} \in \mathcal{G}$, the trajectory $\underline{x}_k(\underline{\tilde{x}})$ originating in $\underline{\tilde{x}}$ remains in \mathcal{G} for all k .

Remark 1. A nonempty subset $\mathcal{S} \subset R^n$ is a *positively invariant set* of the system (S^a) if for any initial state $\underline{x}_o \in \mathcal{S}$, the trajectory $\underline{x}_k(\underline{x}_o) \in \mathcal{S}$ for all k , or equivalently [8] if and only if $\underline{x} \in \mathcal{S}$ implies $A\underline{x} \in \mathcal{S}$. Therefore, it follows that the system (S^a) is constrained stable if and only if it has the set \mathcal{G} as a positively invariant set.

Next, we take into account parametric model uncertainty by extending the concept of constrained stability to a family of systems.

DEFINITION 2. Consider the family of linear discrete-time systems modeled by the difference equation

$$(S^a_{\Delta}) \quad \underline{x}_{k+1} = (A + \Delta)\underline{x}_k$$

where Δ belongs to some perturbation set $\mathcal{D} \subseteq R^{n \times n}$. The system (S^a) is *robustly constrained stable* with respect to the set \mathcal{D} if (S^a_{Δ}) is constrained stable for all perturbation matrices $\Delta \in \mathcal{D}$.

We now restrict the class of constraints allowed in our problem. As it will become apparent later, the introduction of this restriction, termed the *constraint qualification*

hypothesis, while not affecting significantly the number of real-world problems that can be handled by our formalism, introduces more structure into the problem. This additional structure plays a key role in §3 where we derive necessary and sufficient conditions for constrained stability.

2.2. Constraint qualification hypothesis. In this paper, we will limit ourselves to constraints of the form

$$(2) \quad \underline{x} \in \mathcal{G} \subset R^n$$

where \mathcal{G} is a convex, compact, balanced set (i.e., a convex compact set such that $\underline{x} \in \mathcal{G} \Rightarrow \lambda \underline{x} \in \mathcal{G}$ for $|\lambda| \leq 1$ [9]) containing the origin in its interior.

DEFINITION 3. [9]. The *Minkowsky functional* (or gauge) p of a balanced convex set \mathcal{G} containing the origin in its interior is defined by

$$(3) \quad p(\underline{x}) = \inf_{r>0} \left\{ r: \frac{\underline{x}}{r} \in \mathcal{G} \right\}.$$

A well-known result in functional analysis (see, for instance, [9]) establishes that p defines a seminorm in R^n . Furthermore, when \mathcal{G} is compact, this seminorm becomes a norm. In the sequel, we will denote this norm as

$$\|\underline{x}\|_{\mathcal{G}} \triangleq p(\underline{x}).$$

Remark 2. The set \mathcal{G} can be characterized as the unity ball in $\|\cdot\|_{\mathcal{G}}$, i.e., $\mathcal{G} = \{\underline{x}: \|\underline{x}\|_{\mathcal{G}} \leq 1\}$.

With the concepts introduced in this section, we are now ready to give a formal definition to our problem.

2.3. Statement of the problem. Consider the LTI system represented by the following state-space realization:

$$(S) \quad \underline{x}_{k+1} = A\underline{x}_k + B\underline{u}_k$$

subject to the constraint

$$\underline{x}_k \in \mathcal{G} \subset R^n$$

where $\underline{x} \in R^n$ represents the state and $\underline{u} \in R^m$ represents the control input. Then, the basic problems that we address in this paper are the following.

Robust constrained stability analysis problem. Given the nominal system (S) and a linear feedback control law $\underline{u}_k = F\underline{x}_k$, determine if the resulting closed-loop system is constrained stable. If the nominal closed-loop system is constrained stable, determine the maximum allowable level of model uncertainty (in the sense of some previously defined norm) such that the constraints are satisfied for any initial condition $\underline{x} \in \mathcal{G}$.

Linear robust constrained control synthesis problem. Given the system (S), find a *linear* controller such that the resulting closed-loop system is constrained stable and satisfies some additional specifications such as:

- (i) maximum robustness against structured model uncertainty of the form $A = A_o + \Delta$, $\Delta \in \mathcal{D}$;
- (ii) bounds on the control effort of the form $\underline{u}_k \in \Omega \subset R^m$, where Ω is a compact, convex, balanced set containing the origin in its interior.

3. Constrained stability analysis. Consider the system (S^a) and let $\|\cdot\|_{\mathcal{G}}$ denote the operator norm induced in $R^{n \times n}$ by \mathcal{G} , (i.e., $\|A\|_{\mathcal{G}} \triangleq \sup_{\|x\|_{\mathcal{G}}=1} \|Ax\|_{\mathcal{G}}$). From Definition 1 it follows that (S^a) is constrained stable if and only if $\|A\|_{\mathcal{G}} \leq 1$. Moreover, (S^a) is robustly constrained stable with respect to a given set \mathcal{D} if and only if $\|A + \Delta\|_{\mathcal{G}} \leq 1$ for all $\Delta \in \mathcal{D}$. This observation can be used to define a robustness measure in terms of the size of the smallest destabilizing perturbation as follows.

DEFINITION 4. Consider the system (S^a) . The *constrained stability measure* $\varrho_{\mathcal{G}}^{\mathcal{N}}$ is defined as

$$\varrho_{\mathcal{G}}^{\mathcal{N}} \triangleq \begin{cases} 0 & \text{if } \|A\|_{\mathcal{G}} > 1; \\ \max_{\Delta \in \mathcal{D}} \|\Delta\|_{\mathcal{N}} & \text{if } \|A + \Delta\|_{\mathcal{G}} < 1 \ \forall \Delta \in \mathcal{D}; \\ \min_{\Delta \in \mathcal{D}} \{\|\Delta\|_{\mathcal{N}} : \|A + \Delta\|_{\mathcal{G}} = 1\} & \text{otherwise} \end{cases}$$

where $\|\cdot\|_{\mathcal{N}}$ denotes a suitable operator norm defined in \mathcal{D} . In the special case where the induced operator norm $\|\cdot\|_{\mathcal{G}}$ is used in the set \mathcal{D} , we will denote the constrained stability measure as $\varrho_{\mathcal{G}}$.

Remark 3. Let the set $\mathcal{B}\Delta^{\mathcal{N}}$ be the intersection of \mathcal{D} with the origin centered ball of radius $\varrho_{\mathcal{G}}^{\mathcal{N}}$, i.e.,

$$\mathcal{B}\Delta^{\mathcal{N}} = \{\Delta \in \mathcal{D} : \|\Delta\|_{\mathcal{N}} \leq \varrho_{\mathcal{G}}^{\mathcal{N}}\}.$$

Then, from Definition 4 it follows that the family (S_{Δ}^a) is constrained stable for all perturbations $\Delta \in \mathcal{B}\Delta^{\mathcal{N}}$.

Remark 4. Definition 4 is quite general since in principle no conditions are imposed over the set \mathcal{D} . However, in the general case nothing can be stated about the properties of $\varrho_{\mathcal{G}}^{\mathcal{N}}$ which could conceivably be a *noncontinuous* function of A . In the sequel we will show that, under some assumptions that are commonly verified in practice, $\varrho_{\mathcal{G}}^{\mathcal{N}}$ is a *continuous, concave* function of the dynamics matrix A .

THEOREM 1. Assume that the perturbation set \mathcal{D} is a closed cone with vertex at the origin [10], (i.e., $\Delta^o \in \mathcal{D} \iff \alpha \Delta^o \in \mathcal{D}$ for all $0 \leq \alpha$). Then $\varrho_{\mathcal{G}}^{\mathcal{N}}$ is a continuous, concave function of A .

Proof. The proof of the theorem is given in Appendix A.

Remark 5. Note that the class of sets considered in this theorem includes, as a particular case, sets of the form:

$$(4) \quad \mathcal{D} = \left\{ \Delta : \Delta = \sum_1^m \mu_i E_i; \ \mu_i \geq 0, \ E_i \text{ given} \right\}$$

which has been the object of much interest lately ([11]–[13] and references therein).

In the next lemma we introduce a *lower bound* of the constrained stability measure and we show that for *unstructured* perturbations (i.e., the case where $\mathcal{D} \equiv R^{n \times n}$) this lower bound is saturated.

LEMMA 1.

$$(5) \quad \varrho_{\mathcal{G}} \geq 1 - \|A\|_{\mathcal{G}}.$$

Furthermore, for the unstructured perturbation case, i.e., the case where $\mathcal{D} \equiv R^{n \times n}$, condition (5) is saturated.

Proof. The first part of the lemma can be easily proved from Definition 4 and the triangle inequality. The second part follows by noting that for $\Delta^o \triangleq ((1 - \|A\|_{\mathcal{G}})A / \|A\|_{\mathcal{G}})$ (5) is saturated. \square

Remark 6. Note that the results of Lemma 1 can be used to find a lower bound for the constrained robustness measure in the general case when an operator norm different from $\|\cdot\|_{\mathcal{G}}$ is used in the set \mathcal{D} . Since all finite-dimensional matrix norms are equivalent [14], it follows that, given any norm \mathcal{N} in the set \mathcal{D} , there exist a constant c such that $\|\cdot\|_{\mathcal{G}} \leq c\|\cdot\|_{\mathcal{N}}$. Hence $\varrho_{\mathcal{G}}^{\mathcal{N}} \leq (\varrho_{\mathcal{G}}/c)$.

3.1. Quadratic constraints case. In this section we particularize our theoretical results for the special case where the constraint region is an hyperellipsoid. In this case, without loss of generality, we have

$$\mathcal{G} = \{\underline{x}: \underline{x}'P\underline{x} \leq 1, P \in R^{n \times n} \text{ positive definite}\}.$$

Hence

$$\begin{aligned} \|\underline{x}\|_{\mathcal{G}}^2 &= \underline{x}'P\underline{x} \\ \|A\|_{\mathcal{G}}^2 &= \max_{\underline{x}} \left\{ \frac{\underline{x}'A'PA\underline{x}}{\underline{x}'P\underline{x}} \right\} \\ (6) \quad &= \max_{\underline{x}} \left\{ \frac{\underline{x}'L'L^{-1}A'L'LAL^{-1}L\underline{x}}{\underline{x}'L'L\underline{x}} \right\} \\ &= \max_{\|\underline{y}\|_2=1} \|LAL^{-1}\underline{y}\|_2^2 \\ &= \|LAL^{-1}\|_2^2 = \|\tilde{A}\|_2^2 \end{aligned}$$

where $L'L = P$ and $\tilde{A} = LAL^{-1}$. We will show that in this case our approach yields a generalization of the well-known technique of estimating the robustness measure by using quadratic based Lyapunov functions, (see [15] and references therein) by obtaining robustness bounds previously derived in this context.

Example 1 (multilinearly correlated perturbations). In the case of quadratic constraints and multilinearly correlated uncertainty, the lower bound on ϱ given by (5) can be tightened as follows. Assume that the set \mathcal{D} is given by

$$(7) \quad \mathcal{D} = \left\{ \Delta \in R^{n \times n}: L\Delta L^{-1} = U \begin{pmatrix} \tilde{\Delta} \\ 0 \end{pmatrix}, \tilde{\Delta} \in R^{m \times n}, U'U = I_n, L'L = P \right\}.$$

Since the euclidian norm is invariant under multiplications by a unitary matrix, we have

$$\begin{aligned} \|A + \Delta\|_{\mathcal{G}} &= \|L(A + \Delta)L^{-1}\|_2 \\ (8) \quad &= \left\| \tilde{A} + U \begin{pmatrix} \tilde{\Delta} \\ 0 \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} + \begin{pmatrix} \tilde{\Delta} \\ 0 \end{pmatrix} \right\|_2 \\ &= \left\| \begin{pmatrix} A_1 + \tilde{\Delta} \\ A_2 \end{pmatrix} \right\|_2. \end{aligned}$$

A well-known result on matrix dilations establishes [16] that

$$\left\| \begin{pmatrix} X \\ A_2 \end{pmatrix} \right\|_2 \leq 1 \iff \|A_2\|_2 \leq 1 \quad \text{and} \quad X = Y(I - A_2'A_2)^{1/2}, \quad \|Y\|_2 \leq 1;$$

hence it follows that

$$(9) \quad \|A + \Delta\|_{\mathcal{G}} = 1 \iff \|(A_1 + \tilde{\Delta})N\|_2 = 1$$

where:

$$N \triangleq (I - A_2' A_2)^{-1/2}.$$

Finally, by defining $\|\Delta\|_{\mathcal{N}} \triangleq \|\tilde{\Delta}N\|_2$ and using the results of Lemma 1, we get

$$(10) \quad \varrho_{\mathcal{G}}^{\mathcal{N}} = 1 - \|A_1 N\|_2.$$

Remark 7. Note that when $A_2 = 0$ we recover the results of Lemma 1, since in this case $\varrho = 1 - \|A_1\|_2 = 1 - \|A\|_{L'L}$.

Example 2. (unstructured perturbation). In this case, Theorem 2 yields $\varrho_{\mathcal{G}} = 1 - \|A\|_{\mathcal{G}}$ where

$$(11) \quad \|A\|_{\mathcal{G}}^2 = \|A\|_P^2 = \max_{\underline{x}} \left(\frac{\underline{x}' A' P A \underline{x}}{\underline{x}' P \underline{x}} \right).$$

Now consider the case where $\varrho_{\mathcal{G}} > 0$. Then there exists Q positive definite such that

$$(12) \quad A' P A - P = -Q$$

and

$$(13) \quad \|A\|_{\mathcal{G}}^2 = \max_{\underline{x}} \left(1 - \frac{\underline{x}' Q \underline{x}}{\underline{x}' P \underline{x}} \right) \leq 1 - \frac{\sigma_{\min}(Q)}{\sigma_{\max}(P)}.$$

Hence

$$(14) \quad \varrho_{\mathcal{G}} = 1 - \|A\|_{\mathcal{G}} \geq 1 - \left(1 - \frac{\sigma_{\min}(Q)}{\sigma_{\max}(P)} \right)^{1/2}$$

A common technique in state space robust analysis is to obtain robustness bounds from equation (12) [17], [18]. This case can be accommodated by our formalism by recognizing the fact that once P is selected, the system becomes effectively constrained to remain within an hyperellipsoidal region. It has been suggested [17], [18] that good robustness bounds can be obtained from (12) when P is selected such that $Q = I$. In this case our approach yields

$$(15) \quad \varrho_{\mathcal{G}} = 1 - \|A\|_{\mathcal{G}} = 1 - \left(1 - \frac{1}{\sigma_{\max}(P)} \right)^{1/2}$$

which coincides with the robustness bound found by Sezer and Siljak [18].

Example 3 (unstructured perturbation, A semisimple). Consider the case where A is semisimple, i.e.,

$$(16) \quad \begin{aligned} A &= L^{-1} \Lambda L \\ \Lambda &= \text{diag} \left\{ \begin{pmatrix} \sigma_1 & \omega_1 \\ -\omega_1 & \sigma_1 \end{pmatrix}, \dots, \begin{pmatrix} \sigma_p & \omega_p \\ -\omega_p & \sigma_p \end{pmatrix}, \sigma_{p+1}, \dots, \sigma_n \right\}. \end{aligned}$$

Then, the maximum of the stability measure, $\varrho_{\mathcal{G}}$, over all possible positive definite matrices P , is achieved for $P = L'L$.

Proof. The proof follows by noting that $\|A\|_{L'L} = \rho(A)$ where $\rho(\cdot)$ denotes the spectral radius, which is a lower bound for any matrix norm [14]. \square

3.2. Polyhedral constraints. Consider now the case where the region \mathcal{G} is polyhedral, i.e., the case where

$$(17) \quad \mathcal{G} = \{\underline{x}: |G\underline{x}| \leq \underline{\omega}\}$$

where $G \in R^{p \times n}$, $\text{rank}(G) = n$, $\underline{\omega} \in R^p$, $\omega_i > 0$ and the $|\cdot|$ should be interpreted on a component by component sense. Although this case is of practical importance, up to now a technique to estimate the robustness of such systems was unavailable, except perhaps to fit an hyperellipsoidal region within the admissible region and then use some of the bounds available for the quadratic case. Such a technique is clearly inappropriate since it guarantees robust stability *only* in a certain subregion of the region of interest. In this section we show that polyhedral regions fit naturally within our formalism and that in this case $\varrho_{\mathcal{G}}^{\mathcal{N}}$ can be efficiently computed as the minimum of the solution of p linear programming problems.

THEOREM 2. Let $\varrho_i^{\mathcal{N}}$ be the solution of the following optimization problem:

$$(18) \quad \varrho_i^{\mathcal{N}} = \min_{\Delta \in \mathcal{D}} \{\|\Delta\|_{\mathcal{N}}: \|H + \Delta H\|_1^{(i)} \geq 1\}$$

where

$$\begin{aligned} W &= \text{diag}\{w_i\} \\ H &\triangleq W^{-1}GA(G'G)^{-1}G'W \\ \Delta H &\triangleq W^{-1}G\Delta(G'G)^{-1}G'W \end{aligned}$$

and where $\|M\|_1^{(i)}$ indicates the l_1 norm of the i th row of the matrix M . Then

$$(19) \quad \varrho_{\mathcal{G}}^{\mathcal{N}} = \min_{1 \leq i \leq p} \{\varrho_i^{\mathcal{N}}\}.$$

Proof. It is easily shown that

$$(20) \quad \|\underline{x}\|_{\mathcal{G}} = \max_{1 \leq i \leq p} \left\{ \frac{|G\underline{x}|_i}{\omega_i} \right\} = \|W^{-1}G\underline{x}\|_{\infty}.$$

From the definition of H we have that $W^{-1}GA = HW^{-1}G$. Hence

$$\|A\underline{x}\|_{\mathcal{G}} = \|W^{-1}GA\underline{x}\|_{\infty} = \|HW^{-1}G\underline{x}\|_{\infty}$$

and

$$\|A\|_{\mathcal{G}} = \sup_{\|W^{-1}G\underline{x}\|_{\infty}=1} \|A\underline{x}\|_{\mathcal{G}} = \sup_{\|\underline{y}\|_{\infty}=1} \|H\underline{y}\|_{\infty} = \|H\|_{\infty}.$$

Assume that the lemma is false and that there exist $\tilde{\varrho}$ and $\tilde{\Delta}$ such that

$$(21) \quad \|A + \tilde{\Delta}\|_{\mathcal{G}} = 1; \quad \|\tilde{\Delta}\|_{\mathcal{N}} = \tilde{\varrho} < \varrho_{\mathcal{G}}^{\mathcal{N}}.$$

Since $\|A + \tilde{\Delta}\|_{\mathcal{G}} = 1$ there exists i^o such that $\|H + \tilde{\Delta}H\|_1^{(i^o)} = 1$, $\|H + \tilde{\Delta}H\|_1^{(j)} \leq 1$, $j \neq i^o$, but this implies (equation (18)) that $\varrho_{i^o}^{\mathcal{N}} \leq \tilde{\varrho}$, which contradicts (21). \square

Example 4 (unstructured perturbation). Consider the following case:

$$(22) \quad A = \begin{pmatrix} 0.8 & 0.5 \\ -0.0208 & 0.5083 \end{pmatrix} \quad G = \begin{pmatrix} 1.0 & 2.0 \\ -1.5 & 2.0 \end{pmatrix} \quad \underline{\omega} = \begin{pmatrix} 5.0 \\ 10.0 \end{pmatrix}.$$

Then, from the definition of H , we have that

$$(23) \quad H = \begin{pmatrix} 0.7583 & 0.0 \\ -0.2083 & 0.55 \end{pmatrix}, \quad \|A\|_{\mathcal{G}} = 0.7583$$

and, from Theorem 2,

$$(24) \quad \varrho_i = \min_{\|\Delta\|_{\mathcal{G}}} \left\{ \|\Delta\|_{\mathcal{G}} : \sum_{j=1}^2 |H + \Delta|_{ij} = 1 \right\} \quad i = 1, 2.$$

Casting the problems (24) into a linear programming form and solving we have that

$$\varrho_1 = 0.2417, \quad \varrho_2 = 0.2417 \quad \text{and} \quad \varrho_{\mathcal{G}} = \min_{1 \leq i \leq 2} \varrho_i = 0.2417.$$

Note that, in this case, $\varrho_{\mathcal{G}} = 1 - \|A\|_{\mathcal{G}} = 0.2417$ as shown in Lemma 1.

4. Application to robust controllers design. Consider the *linear robust constrained control synthesis problem* introduced in §2.3. Let $p_{\Omega}(u)$ be the Minkowsky gauge for the set Ω and denote by $\|\cdot\|_{\Omega}$ the corresponding norm induced in R^m . It follows that, given a feedback control law of the form $u_k = Fx_k$, the control bounds are satisfied if and only if

$$\|F\|_{\mathcal{G}, \Omega} \triangleq \sup_{\|x\|_{\mathcal{G}} \leq 1} \|Fx\|_{\Omega} \leq 1.$$

Hence, a full state feedback matrix F that solves the synthesis problem can be found solving the following optimization problem:

$$(25) \quad \max_F \{ \varrho_{\mathcal{G}}^{\mathcal{N}}(F) \}$$

subject to

$$(26) \quad \begin{aligned} \varrho_{\mathcal{G}}^{\mathcal{N}}(F) &\triangleq \min_{\Delta \in \mathcal{D}} \{ \|\Delta\|_{\mathcal{N}} : \|A + BF + \Delta\|_{\mathcal{G}} = 1 \} \\ \|F\|_{\mathcal{G}, \Omega} &\leq 1. \end{aligned}$$

Since from Theorem 1, $\varrho_{\mathcal{G}}^{\mathcal{N}}(F)$ is a concave function, and since $\|F\|_{\mathcal{G}, \Omega} \leq 1$ is a convex constraint, it follows that (25) has a global optimum. Hence, the problem of finding the *maximally* robust controller leads to convex, albeit nondifferentiable, optimization problems, which can be solved using a number of techniques [19]. In the remainder of this section, we give several design examples using the proposed technique.

Example 5. Consider the following system:

$$(27) \quad \begin{aligned} A &= \begin{pmatrix} 0 & 1 \\ 0.505 & -0.51 \end{pmatrix} \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ \mathcal{G} &= \{x : \|x\|_2 \leq 1\}. \end{aligned}$$

The open-loop system has poles at $s_1 = 0.5$ and $s_2 = -1.01$. Assume that the perturbation set is such that it changes the position of the poles while maintaining constant their sum, i.e.,

$$(28) \quad \mathcal{D} = \left\{ \Delta : \Delta = \mu E, \quad E \triangleq \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mu \in \mathbb{R} \right\}.$$

Note that $\|E\|_2 = 1$; hence $\|\Delta\|_2 = |\mu|$.

In this case, the solution to the unconstrained maximally robust control problem can be computed by solving a matrix dilation problem [16]. Rewrite the dynamics matrix as

$$A = \begin{pmatrix} x_1 & x_2 \\ a_1 & a_2 \end{pmatrix},$$

where x_i denote elements that can be modified using state feedback. Since matrix dilations are norm increasing we have that

$$(29) \quad \begin{aligned} \|A + \mu E\|_2 &\geq \max \{ \| \begin{pmatrix} a_1 & a_2 + \mu \end{pmatrix} \|_2 \} \\ &= \sqrt{a_1^2 + (a_2 + \mu)^2}. \end{aligned}$$

Now define:

$$(30) \quad \begin{aligned} \mu^0 &= \operatorname{argmin} \{ |\mu|, \mu \in \mathbb{R}: a_1^2 + (a_2 + \mu)^2 = 1 \} \\ &= \sqrt{(1 - a_1^2)} - |a_2|. \end{aligned}$$

From (29) and (30) it follows that $\|A + \mu^0 E\|_2 \geq 1$ which implies that $\varrho_2(F) \leq \mu^0$ for all F . Furthermore, from the definition of μ^0 it follows that if F is selected such that $x_1 = x_2 = 0$, then $\varrho_2(F) = \mu^0$. Hence, this choice of F yields the solution to the unconstrained problem. In this particular example we have

$$(31) \quad F^o = \begin{pmatrix} 0 & 1 \end{pmatrix}, \quad \varrho_2 = 0.3531.$$

Now consider a feedback matrix F and let A_{cl} be the corresponding *closed-loop* matrix, i.e.,

$$(32) \quad A_{cl} = A + BF = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} + \mu \end{pmatrix}.$$

The corresponding value of the robustness measure can be computed using standard results on matrix dilations [16] as follows: The set Υ of numbers μ such that $\|A_{cl}\|_2 \leq 1$ can be parametrized as

$$(33) \quad \Upsilon = \{ \mu: \mu = -a_{22} - ya_{11}z + (1 - y^2)^{1/2}w(1 - z^2)^{1/2} \}$$

where

$$(34) \quad \begin{aligned} y &= \frac{a_{21}}{(1 - a_{11}^2)^{1/2}} \\ z &= \frac{a_{12}}{(1 - a_{11}^2)^{1/2}} \\ w &\in \mathbb{R}, |w| \leq 1. \end{aligned}$$

From (33) it follows that the constrained stability margin of A_{cl} is given by

$$\varrho_2(F) = |a_{22} + ya_{11}z - (1 - y^2)^{1/2}(1 - z^2)^{1/2}\operatorname{sign}(a_{22} + ya_{11}z)|.$$

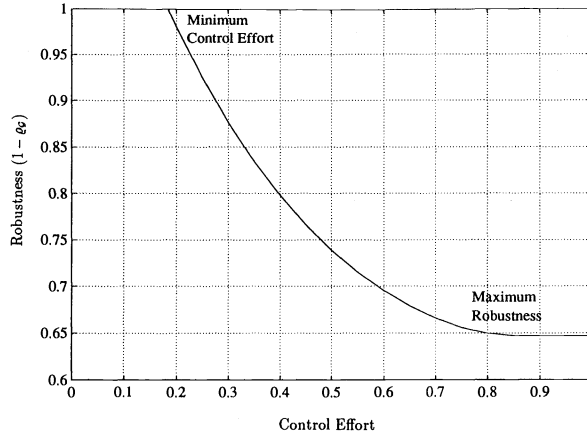


FIG 1. Robustness vs. control effort for Example 5.

Figure 1 shows $\rho_2(F)$ versus $\|F\|_2$, the norm of the solution to (25). For $\|F\|_2 = 1$, we recover the unconstrained solution, for $\|F\|_2 = 0.1850$, we get the minimum control effort capable of stabilizing (in the constrained sense) the nominal system. Note the trade-off between control effort and robustness. In particular, there exist a region where the curve is flat, i.e., the control effort can be reduced while essentially maintaining the same robustness obtained with a “maximum robustness” type design.

Example 6 (polyhedral constraints, unstructured perturbation). Consider the following system:

$$\begin{aligned}
 A &= \begin{pmatrix} 0.8 & 0.5 \\ -0.4 & 1.2 \end{pmatrix} & B &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
 G &= \begin{pmatrix} 1.0 & 2.0 \\ -1.5 & 2.0 \end{pmatrix} & \underline{\omega} &= \begin{pmatrix} 5.0 \\ 10.0 \end{pmatrix} & \Omega &= \{u: |u| \leq \gamma\}.
 \end{aligned}
 \tag{35}$$

Since the constraint sets \mathcal{G} and Ω are polyhedral, the synthesis problem can be cast in the following format:

$$\min_F \epsilon$$

subject to

$$\begin{aligned}
 \|A + BF\|_{\mathcal{G}} &\leq \epsilon \\
 \|F\|_{\mathcal{G}, \Omega} &\leq 1,
 \end{aligned}$$

which can be transformed into an LP problem and solved using the simplex method. Note that a similar design algorithm was proposed by Vassilaki, Hennet, and Bitsoris [20], although in their case the goal was to find admissible linear controllers for systems under polyhedral constraints without taking into account robustness considerations. Figure 2 shows the constrained robustness measure versus γ , the bound on the control effort. Note that the minimum control effort required to stabilize the system is $\gamma = 2.6$.

5. Conclusions. Most realistic control problems involve both some type of time-domain constraints and a certain degree of model uncertainty. However, the majority

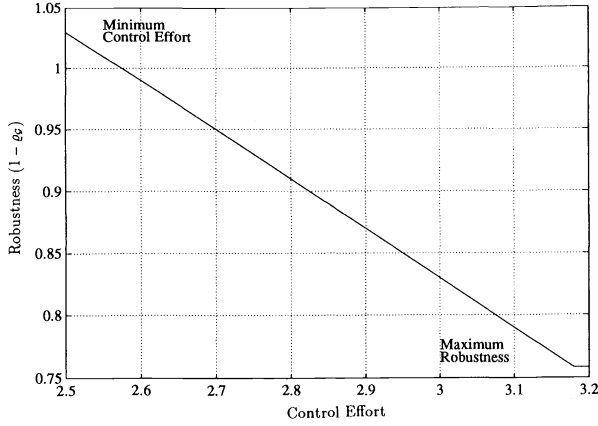


FIG 2. Robustness vs. control effort for example 6.

of control design methods currently available focus only on one aspect of the problem. Following the spirit of [6], [7], in this paper we proposed to approach time-domain constraints using an operator norm induced by the constraints to assess the stability properties of a family of systems. Specifically, in §2 we introduced a robustness measure that indicates how well the family of systems under consideration satisfies a given set of constraints. In §3 we explored the properties of this robustness measure for the case of additive parametric model uncertainty and we showed that our formalism provides a generalization of the well-known technique of estimating robustness bounds from the solution of a Lyapunov equation. We then proposed, in §4, a synthesis procedure for fixed-order controllers, based upon maximization of the robustness measure subject to additional performance constraints such as bounds on the control effort. There we showed that the proposed design procedure leads to a convex optimization problem. We believe that the results presented here will provide a valuable new approach to the problems of robust controllers analysis and design for linear systems. Furthermore, since our approach is based purely upon time-domain analysis, we have reason to believe the theory could be extended to encompass nonlinear systems in a much more direct fashion than other currently used techniques.

Perhaps the more severe limitation of the theory in its present form arises from the fact that the incorporation of additional performance constraints, of the form of a bound on the norm of a relevant transfer function, results in nonconvex optimization problems. We are currently looking into a solution to this problem by using an observer-based parametrization of all stabilizing controllers. It is expected that this formulation will be able to handle more general performance constraints as well as dynamic uncertainty, at the price of resulting in higher-order controllers.

Appendix A. Proof of Theorem 1. We begin by introducing two preliminary results.

LEMMA 2. Consider the system (S^a) . Assume that the perturbation set \mathcal{D} is a closed cone with vertex at the origin [10], i.e., $\Delta^o \in \mathcal{D} \iff \alpha \Delta^o \in \mathcal{D}$ for all $0 \leq \alpha$ and that (S^a) is constraint stable (i.e., $\|A\|_G < 1$). Let

$$(A1) \quad \Delta^o = \operatorname{argmin}_{\Delta \in \mathcal{D}} \{ \|\Delta\|_{\mathcal{N}} : \|A + \Delta\|_G = 1 \}$$

and consider a sequence $A^i \rightarrow A$ such that $\|A^i\|_G < 1$. Finally, define the sequence λ^i

as

$$(A2) \quad \lambda^i = \min_{\lambda \in \mathbb{R}^+} \{ \lambda : \|A^i + \lambda \Delta^o\|_{\mathcal{G}} = 1 \}.$$

Then the sequence λ^i has an accumulation point at one.

Proof. Since $\|A^i\|_{\mathcal{G}} < 1$ and since \mathcal{D} is a closed cone, it follows that λ^i is well defined. Furthermore, from (A2) it follows that

$$(A3) \quad \lambda^i \leq \frac{1 + \|A^i\|_{\mathcal{G}}}{\|\Delta^o\|_{\mathcal{G}}} \leq \frac{2}{\|\Delta^o\|_{\mathcal{G}}}.$$

Hence, from Bolzano–Weierstrass’ theorem [21] it follows that λ^i has an accumulation point $\tilde{\lambda}$ and that there exists a subsequence $\tilde{\lambda}^i \rightarrow \tilde{\lambda}$. Hence

$$\|A^i + \tilde{\lambda}^i \Delta^o\|_{\mathcal{G}} = 1,$$

and since $A^i \rightarrow A$ then

$$(A4) \quad \|A + \tilde{\lambda} \Delta^o\|_{\mathcal{G}} = 1.$$

Assume that $\tilde{\lambda} < 1$ and let $\hat{\Delta} \triangleq \tilde{\lambda} \Delta^o$. Then $\|\hat{\Delta}\|_{\mathcal{N}} < \|\Delta^o\|_{\mathcal{N}}$, $\|A + \hat{\Delta}\|_{\mathcal{G}} = 1$, and $\hat{\Delta} \in \mathcal{D}$ (since \mathcal{D} is a cone) which contradicts (A1). Now assume that $\tilde{\lambda} > 1$. Then, for i large enough, $\tilde{\lambda}^i > 1$, which together with (A2) implies that

$$(A5) \quad \|A^i + \Delta^o\|_{\mathcal{G}} < 1$$

and hence:

$$(A6) \quad \|A + \Delta^o\|_{\mathcal{G}} < 1,$$

which contradicts (A1). Therefore $\tilde{\lambda} = 1$. \square

LEMMA 3. Let $\rho_1 > 0, \rho_2 > 0$ and $0 \leq \lambda \leq 1$ be given numbers and assume that \mathcal{D} is a cone with vertex at the origin. Consider the following sets:

$$(A7) \quad \begin{aligned} \rho_1 B\Delta &= \{\Delta \in \mathcal{D} : \|\Delta\|_{\mathcal{N}} \leq \rho_1\} \\ \rho_2 B\Delta &= \{\Delta \in \mathcal{D} : \|\Delta\|_{\mathcal{N}} \leq \rho_2\} \\ \rho B\Delta &= \{\Delta \in \mathcal{D} : \|\Delta\|_{\mathcal{N}} \leq \rho \triangleq \lambda \rho_1 + (1 - \lambda) \rho_2\}. \end{aligned}$$

Then

$$\rho B\Delta \subseteq \lambda \rho_1 B\Delta + (1 - \lambda) \rho_2 B\Delta.$$

Proof. Consider any $\Delta^o \in \rho B\Delta$. Then

$$(A8) \quad \begin{aligned} \Delta^o &= \frac{\|\Delta^o\|_{\mathcal{N}}}{\rho} \left[\frac{\rho \Delta^o}{\|\Delta^o\|_{\mathcal{N}}} \right] \\ &= \frac{\|\Delta^o\|_{\mathcal{N}}}{\rho} \left[\lambda \rho_1 \frac{\Delta^o}{\|\Delta^o\|_{\mathcal{N}}} + (1 - \lambda) \rho_2 \frac{\Delta^o}{\|\Delta^o\|_{\mathcal{N}}} \right] \\ &= [\lambda \Delta_1 + (1 - \lambda) \Delta_2] \end{aligned}$$

where

$$\begin{aligned}
 \Delta_1 &= \alpha \rho_1 \frac{\Delta^o}{\|\Delta^o\|_{\mathcal{N}}} \\
 \Delta_2 &= \alpha \rho_2 \frac{\Delta^o}{\|\Delta^o\|_{\mathcal{N}}} \\
 \alpha &= \frac{\|\Delta^o\|_{\mathcal{N}}}{\rho} \leq 1.
 \end{aligned}
 \tag{A9}$$

The proof is completed by noting that from (A9) and the hypothesis it follows that $\Delta_1 \in \rho_1 B\Delta$ and $\Delta_2 \in \rho_2 B\Delta$. \square

Proof of Theorem 1. Assume that $\varrho_{\mathcal{G}}^{\mathcal{N}}$ is not continuous. Then, given $\epsilon > 0$, for every $\delta > 0$ there exist A_δ such that $\|A_\delta - A\|_{\mathcal{G}} \leq \delta$ and $|\varrho_{\mathcal{G}}^{\mathcal{N}}(A_\delta) - \varrho_{\mathcal{G}}^{\mathcal{N}}| > \epsilon$. Hence there exist a sequence $A^i \rightarrow A$ such that $\varrho_{\mathcal{G}}^{\mathcal{N}^i} \not\rightarrow \varrho_{\mathcal{G}}^{\mathcal{N}}$. Furthermore, it is easily seen that the sequence $\varrho_{\mathcal{G}}^{\mathcal{N}^i}$ is bounded and therefore it contains a convergent subsequence. It follows that there exists a sequence $A^i \rightarrow A$ such that $\varrho_{\mathcal{G}}^{\mathcal{N}^i} \rightarrow \tilde{\varrho} \neq \varrho_{\mathcal{G}}^{\mathcal{N}}$. Let

$$\Delta^i = \operatorname{argmin}_{\Delta \in \mathcal{D}} \{\|\Delta\|_{\mathcal{N}} : \|A^i + \Delta\|_{\mathcal{G}} = 1\}.
 \tag{A10}$$

From (A10) it follows that $\|\Delta^i\|_{\mathcal{G}} \leq 1 + \|A^i\|_{\mathcal{G}}$. It follows then that the sequence Δ^i is bounded and therefore, since $R^{n \times n}$ with a finite-dimensional matrix norm is complete and since \mathcal{D} is a closed set, it has an accumulation point $\tilde{\Delta}$ (Bolzano–Weierstrass) and a convergent subsequence $\tilde{\Delta}^i \rightarrow \tilde{\Delta}$ such that $\|A + \tilde{\Delta}\|_{\mathcal{G}} = 1$. Furthermore, from the definition of Δ^o it follows that

$$\tilde{\varrho} = \|\tilde{\Delta}\|_{\mathcal{N}} > \|\Delta^o\|_{\mathcal{N}} = \varrho_{\mathcal{G}}^{\mathcal{N}}.
 \tag{A11}$$

Hence, for i large enough,

$$\|\tilde{\Delta}^i\|_{\mathcal{N}} > \|\Delta^o\|_{\mathcal{N}}.
 \tag{A12}$$

Applying Lemma 3, we have that there exists a sequence $\lambda^i \rightarrow 1$ such that

$$\lambda^i = \min_{\lambda \in \mathbb{R}^+} \{\lambda : \|A^i + \lambda \Delta^o\|_{\mathcal{G}} = 1\}.
 \tag{A13}$$

From (A12), and since $\lambda^i \rightarrow 1$, it follows that for i large enough

$$\begin{aligned}
 \|\lambda^i \Delta^o\|_{\mathcal{N}} &< \|\tilde{\Delta}^i\|_{\mathcal{N}} \\
 \|A^i + \lambda^i \Delta^o\|_{\mathcal{G}} &= 1
 \end{aligned}
 \tag{A14}$$

and, since \mathcal{D} is a cone, $\lambda^i \Delta^o \in \mathcal{D}$, which contradicts (A10). The proof is completed by noting that since all finite-dimensional matrix norms are equivalent [14]; then continuity in the $\|\cdot\|_{\mathcal{G}}$ norm implies continuity in any other norm defined over $R^{n \times n}$. \square

To prove concavity, start by considering a convex linear combination $A = \lambda A_1 + (1 - \lambda)A_2$, $\lambda \leq 1$ of given matrices A_1 and A_2 . Then, from Lemma 4 it follows that

$$\begin{aligned}
 \max_{\Delta \in \rho B\Delta} \|A + \Delta\|_{\mathcal{G}} &\leq \max_{\substack{\Delta_1 \in \rho_1 B\Delta \\ \Delta_2 \in \rho_2 B\Delta}} \|\lambda(A_1 + \Delta_1) + (1 - \lambda)(A_2 + \Delta_2)\|_{\mathcal{G}} \\
 &\leq \lambda \max_{\Delta_1 \in \rho_1 B\Delta} \|A_1 + \Delta_1\|_{\mathcal{G}} + (1 - \lambda) \max_{\Delta_2 \in \rho_2 B\Delta} \|A_2 + \Delta_2\|_{\mathcal{G}}.
 \end{aligned}
 \tag{A15}$$

Now consider the case where $\rho_1 = \varrho_G^N(A_1)$ and $\rho_2 = \varrho_G^N(A_2)$. Then it follows from the definition of ϱ_G^N that both maximizations in the right-hand side of (A15) yield one and therefore:

$$(A16) \quad \max_{\Delta \in \rho B \Delta} \|A + \Delta\|_G \leq 1.$$

Hence, from the definition of ϱ_G^N ,

$$\varrho_G^N[\lambda A_1 + (1 - \lambda)A_2] \geq \varrho = \lambda \varrho_G^N(A_1) + (1 - \lambda) \varrho_G^N(A_2). \quad \square$$

REFERENCES

- [1] M. SZNAIER, *Suboptimal feedback control of constrained linear systems*, Ph.D. Dissertation, University of Washington, 1989.
- [2] S. BOYD, V. BALAKRISHNAN, C. H. BARRAT, N. M. KHRAISHI, X. LI, D. G. MEYER, AND S. NORMAN, *A new CAD method and associated architectures for linear controllers*, IEEE Trans. Automat. Control, 33 (1988), pp. 268–283.
- [3] E. POLAK AND S. SALCUDEAN, *On the design of linear multivariable feedback systems via constrained nondifferentiable optimization in H_∞ spaces*, IEEE Trans. Automat. Control, 34 (1989), pp. 268–276.
- [4] J. W. HELTON AND A. SIDERIS, *Frequency Response Algorithms for H_∞ Optimization With Time Domain Constraints*, IEEE Trans. Automat. Control, 34 (1989), pp. 427–434.
- [5] A. SIDERIS AND H. ROTSTEIN, *H_∞ Optimization with time domain constraints over a finite horizon*, Proc. 29th IEEE CDC, Hawaii, Dec. 5–7, 1990, pp. 1802–1807.
- [6] M. SZNAIER, *Norm based robust control of constrained discrete time linear systems*, Proc. 29th IEEE CDC, Hawaii, Dec. 5–7, 1990, pp. 1925–1930.
- [7] M. SZNAIER AND A. SIDERIS, *Norm based optimally robust control of constrained discrete time linear systems*, Proc. 1991 ACC, Boston, MA, June 23–25, 1991, pp. 2710–2715.
- [8] J. P. LASALLE, *The stability and control of discrete processes*, in Applied Mathematics Series, Vol. 62, Springer-Verlag, Berlin, New York, 1986.
- [9] J. B. CONWAY, *A course in functional analysis*, in Graduate Texts in Mathematics, Vol. 96, Springer-Verlag, Berlin, New York, 1990.
- [10] D. G. LUENBERGER, *Optimization by vector space methods*, John Wiley, New York, 1969.
- [11] R. K. YEDAVALLI, *Improved measures of stability robustness for linear state space models*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 557–579.
- [12] K. ZHOU AND P. P. KHARGONEKAR, *Stability robustness bounds for linear state-space models with structured uncertainty*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 621–623.
- [13] L. H. KEEL, S. P. BHATTACHARYYA, AND J. W. HOWZE, *Robust control with structured perturbations*, IEEE Trans. Automat. Control, 33 (1988), pp. 68–77.
- [14] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, London, 1985.
- [15] D. D. SILJAK, *Parameter space methods for robust control design: A guided tour*, IEEE Trans. Automat. Control, 34 (1989), pp. 674–687.
- [16] J. C. DOYLE, *Lecture notes in advances in multivariable control*, ONR/Honeywell Workshop, Minneapolis, 1984.
- [17] R. V. PATEL AND M. TODA, *Quantitative measures of robustness for multivariable systems*, Tech. Report TP-8A, Proc. Joint Automat. Contr. Conf., San Francisco, CA, 1980.
- [18] M. E. SEZER AND D. D. SILJAK, *Robust stability of discrete systems*, Internat. J. Control, 48 (1988), pp. 2055–2063.
- [19] F. H. CLARKE, *Optimization and nonsmooth analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley, New York, 1983.
- [20] M. VASSILAKI, J. C. HENNET, AND G. BITSORIS, *Feedback control of discrete-time systems under state and control constraints*, Internat. J. Control, 47 (1988), pp. 1727–1735.
- [21] A. NAYLOR AND G. R. SELL, *Linear operator theory in engineering and science*, in Applied Mathematical Sciences, Vol. 40, Springer-Verlag, Berlin, New York, 1982.

THE FOUR-BLOCK MODEL MATCHING PROBLEM IN l^1 AND INFINITE-DIMENSIONAL LINEAR PROGRAMMING*

OLOF J. STAFFANS†

Abstract. The purpose of this work is fourfold. First the reader is introduced to the present state of the theory behind the solution of the general model matching problem in l^1 . This solution is based on the fact that the model matching problem can be recast as an infinite-dimensional linear programming problem. However, the transformation into linear programming form is highly nonunique, which leads to the second question of discussion, namely, how to find a “good” linear programming formulation. It is shown that the presently available formulations contain some redundancies that limit the applicability of the theory and lead to linear programming systems containing unnecessary degeneracies. The third object of this work is to study problems related to the computational complexity of two different approximation methods for the solution of the infinite-dimensional linear programming systems. Both of these methods are needed in order to get two-sided error bounds on the cutoff error. One complication is that even after the extra redundancies that were mentioned above have been removed, there are certain multi-output problems that contain a massive intrinsic degeneracy. Finally, the convergence properties of the two solution schemes are investigated, and it is explained which properties of the original linear programming formulation are needed for different types of convergence.

Key words. infinite-dimensional linear programming, l^1 -optimal control

AMS(MOS) subject classifications. 93C15, 93C45, 93C55, 93D15, 93B40, 90C05, 90C48

1. Introduction. The standard discrete model matching problem in control theory can be formulated as follows: Three different sequences of matrices with real entries, $\{H(k)\}_{k=0}^\infty$, $\{U(k)\}_{k=0}^\infty$, and $\{V(k)\}_{k=0}^\infty$ are given. The dimensions of H , U , and V are $m \times n$, $m \times p$, and $q \times n$, respectively. The problem is to find a sequence of matrices $\{Q(k)\}_{k=0}^\infty$, of dimension $p \times q$, in such a way that the norm of

$$(1) \quad \Phi = H - K = H - U * Q * V$$

is as small as possible. Here $K = U * Q * V$ is the convolution of U , Q , and V , given by

$$(U * Q * V)(k) = \sum_{i=0}^k U(k-i) \sum_{j=0}^i Q(i-j)V(j), \quad k \geq 0.$$

This problem arises in several different situations; see, e.g., the discussion in [6].

The norm of Φ that we want to minimize is the operator norm of Φ , regarded as an operator from one space to another. Let X be the space of sequences of n -dimensional vectors $\{x(k)\}_{k=0}^\infty$, and let Y be the space of sequences of m -dimensional vectors $\{y(k)\}_{k=0}^\infty$. Then Φ induces a convolution operator from X into Y defined by

$$y(k) = (\Phi * x)(k) = \sum_{i=0}^k \Phi(k-i)x(i).$$

* Received by the editors December 26, 1990; accepted for publication (in revised form) March 5, 1992.

† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland, and Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0123. This research was supported by the Academy of Finland and the National Science Foundation under grants 222/5211/88 and INT-881331.

If we give both X and Y the l^2 -norm (the energy norm), then the operator norm of Φ , i.e., the norm

$$\|\Phi\| = \sup_{\|x\| \leq 1} \|\Phi * x\|,$$

is given by the H^∞ norm of Φ , and in this case the problem is referred to as the H^∞ minimization problem. If we instead give both X and Y the l^∞ -norm (the supremum norm), then the norm of $\Phi = \{\Phi_{ij}(k)\}_{k=0}^\infty$ is given by

$$(2) \quad \|\Phi\|_{l^1} = \max_{i \in \{1, \dots, m\}} \sum_{j=1}^n \sum_{k=0}^\infty |\Phi_{ij}(k)|.$$

This case is usually referred to as the l^1 -minimization problem, and this is the one that we shall consider here. The standard assumptions on H , U , and V are that they belong to l^1 , i.e., their l^1 -norms defined as in (2) (with m and n replaced by the appropriate dimensions) are finite. Also the free parameter Q is required to belong to l^1 . Thus, we want to solve the following problem:

$$(OPT) \quad \begin{array}{l} \text{minimize } \|\Phi\|_{l^1} \text{ over all } \Phi \text{ of the form } \Phi = H - U * Q * V \text{ where } Q \in l^1 \\ \text{and } \|\Phi\|_{l^1} \text{ is given by (2).} \end{array}$$

For any l^1 -function Φ we define the \mathcal{Z} -transform $\hat{\Phi}$ of Φ by

$$\hat{\Phi}(z) = \sum_{k=0}^\infty \Phi(k) z^k.$$

Throughout we denote \mathcal{Z} -transforms as above, i.e., a hat on a symbol representing a sequence in l^1 stands for the \mathcal{Z} -transform of this sequence. The \mathcal{Z} -transforms are analytic in the open unit disk D , and continuous on the closed unit disk \overline{D} . Moreover, the \mathcal{Z} -transform converts convolutions into pointwise multiplications. Thus, the \mathcal{Z} -transform applied to (1) gives

$$(3) \quad \hat{\Phi}(z) = \hat{H}(z) - \hat{K}(z) = \hat{H}(z) - \hat{U}(z)\hat{Q}(z)\hat{V}(z), \quad z \in \overline{D}.$$

Frequently we assume that the matrix functions \hat{H} , \hat{U} , and \hat{V} are (stable) rational functions. In the approach that we use this is not important; it is enough to assume that they belong to l^1 , i.e., they are \mathcal{Z} -transforms of l^1 -sequences. Observe, however, that our \mathcal{Z} -transforms have entries that are real for real z .

Throughout we shall make the following assumption.

Assumption 1.1. The rank of the matrices $\hat{U}(z)$ and $\hat{V}(z)$ are constant on the unit circle $|z| = 1$.

It is known that without this assumption the problem that we consider here does not, in general, have a minimizer Φ ; some examples with boundary zeros where no minimizers exist are given in [17].

The outline of this paper is as follows. In §2 we show that we may, without loss of generality, assume that $\hat{U}(z)$ has full column rank and $\hat{V}(z)$ has full row rank for all but finitely many $z \in D$. Let us assume that this is the case. Then

$$m \geq p \quad \text{and} \quad n \geq q.$$

If $m = p$ and $n = q$ then the problem is called a *one-block problem*. If one of the inequalities above is strict but not the other it is called a *two-block problem*. If both inequalities are strict it is a *four-block problem*. We use the common name *multiblock problem* for the two-block and four-block problems.

In §§3 and 4 we describe the set of all functions $\hat{K} \in \hat{l}^1$ that are of the form $\hat{K} = \hat{U}\hat{Q}\hat{V}$ for some $\hat{Q} \in \hat{l}^1$, i.e., we give a frequency domain description of the range of the operator $\hat{Q} \mapsto \hat{U}\hat{Q}\hat{V}$. Our formulation differs from the one in [10] in two respects: we have removed the redundancies appearing in the equations in [10], as well as one extra nonstandard hypothesis that limits the applicability of the results of [10] to the four-block case. In particular, in the one-block case we are able to identify the exact number of equations that are needed to describe the problem.

In §§5 and 6 we construct an operator whose null-space equals the range of the mapping $Q \mapsto U * Q * V$, and show how we may reformulate the problem as an infinite-dimensional linear programming problem in l^1 . We also present the corresponding dual infinite-dimensional linear programming problems in l^∞ and c^0 . These formulations are quite different in the one-block case and in the multiblock case. In the one-block case only finitely many constraints are needed, and the problem is semifinite (infinitely many variables, but only finitely many constraints). In the multiblock case there are infinitely many variables and infinitely many constraints.

In §7 we describe the standard alignment conditions between an optimal solution of the primal system and an optimal solution of the dual system.

Two different methods to solve the infinite-dimensional linear programming problems are described in §8. The solution of the one-block problem is quite simple compared to the solution of the multiblock problem. In [5] and [10] a solution method for the general case is described, which is based on an iterative procedure where we restrict the number of variables that is allowed to be nonzero, and gradually increase this number. An alternative approach was presented in [3], and independently in [14], where we instead ignore all but finitely many of the constraints, and gradually add more and more constraints. For the one-block problem the two methods are equivalent, and they both give the optimal solution in a finite number of iterations. However, when they are applied to the multiblock problem they give different results, one of them giving upper bounds on the optimum and the other giving lower bounds. As it was pointed out in [3] and [14], when the two methods are combined we get both upper and lower bounds, and we can compute the solution to within any specified accuracy.

The next three sections are devoted to a study of some computational aspects of the two solution methods, and to their convergence properties. The alternative approach mentioned above is especially interesting because it seems to work quite well in many cases, but there is a problem related to the computational complexity of this method. This problem is discussed in §9, and some solutions are proposed.

For the solutions mentioned to work it is crucial that the solution methods have certain convergence properties; it is not enough that the optima of the truncated problems converge to the optimum of the full problem. In addition some of the variables used in the formulation of the problem must converge. We take a closer look at the convergence properties of the two different solution methods in §§10 and 11.

The alternative approach has another drawback: it does not directly produce suboptimal solutions of the original problem. Still, as we show in §12, it is, in general, also possible to find suboptimal solutions in this case.

We shall use the following notation (some of this notation was already introduced

and is similar to that in, e.g., [10]).

$\langle x, x^* \rangle$	The value of the bounded functional x^* evaluated at x .
X^*	The dual of the Banach space X , with norm $\ x^*\ = \sup_{\ x\ \leq 1} \langle x, x^* \rangle$.
S^\perp	The (right) annihilator of $S \subset X$; $S^\perp = \{x^* \in X^* \mid \langle x, x^* \rangle = 0 \text{ for all } x \in S\}$. This set is weak* closed in X^* . Note that $S^\perp \subset X^*$.
${}^\perp S$	The (left) annihilator of $S \subset X^*$; ${}^\perp S = \{x \in X \mid \langle x, x^* \rangle = 0 \text{ for all } x^* \in S\}$. This set is (weakly and strongly) closed in X . Note that ${}^\perp S \subset X$.
$l_{m \times n}^1$	The set of summable $m \times n$ matrix valued sequences H , with norm $\ H\ _{l^1} = \max_{i \in \{1, \dots, m\}} \sum_{j=1}^n \sum_{k=0}^\infty H_{ij}(k) $. This is the dual of $c_{m \times n}^0$, with duality mapping $\langle x, x^* \rangle = \sum_{i,j,k} x_{ij}(k) x_{ij}^*(k)$.
$l_{m \times n}^\infty$	The set of bounded $m \times n$ matrix valued sequences H , with norm $\ H\ _{l^\infty} = \sum_{i=1}^m \sup_{1 \leq j \leq n, k \geq 0} H_{ij}(k) $. This is the dual of $l_{m \times n}^1$, with duality mapping $\langle x, x^* \rangle = \sum_{i,j,k} x_{ij}(k) x_{ij}^*(k)$.
$c_{m \times n}^0$	The closed subset of $l_{m \times n}^\infty$ of sequences H satisfying $ H(k) \rightarrow 0$ as $k \rightarrow \infty$.
D, \overline{D}	The open and closed unit disks in the complex plane, respectively.
\hat{H}	The \mathcal{Z} -transform of a function $H \in l_{m \times n}^1$; $\hat{H}(z) = \sum_{k=0}^\infty H(k) z^k$. It is analytic in D and continuous on \overline{D} .
$\hat{l}_{m \times n}^1$	The space of all \mathcal{Z} -transforms of functions in $l_{m \times n}^1$.
A^*	The adjoint of a continuous operator A mapping X into Y ; it maps Y^* into X^* .
$\mathcal{D}(A)$	Domain of the operator A .
$\mathcal{R}(A)$	Range of the operator A .

Recall that $\mathcal{N}(A) = {}^\perp \mathcal{R}(A^*)$, that $\mathcal{N}(A^*) = \mathcal{R}(A)^\perp$, that $\mathcal{N}(A)^\perp$ is the weak* closure of $\mathcal{R}(A^*)$, and that ${}^\perp \mathcal{N}(A^*)$ is the closure of $\mathcal{R}(A)$; see, e.g., [12, pp. 90–94].

2. Reduction of the dimension of the free parameter. We claim that it is almost always possible to assume, without loss of generality, that $\hat{U}(z)$ has full column rank and $\hat{V}(z)$ has full row rank for all but finitely many $z \in D$. To show this we decompose \hat{U} and \hat{V} into their Smith forms

$$(4) \quad \hat{U}(z) = \hat{L}_U(z) \hat{M}_U(z) \hat{R}_U(z), \quad \hat{V}(z) = \hat{L}_V(z) \hat{M}_V(z) \hat{R}_V(z).$$

Here \hat{L}_U , \hat{R}_U , \hat{L}_V , and \hat{R}_V are square matrices of the appropriate dimensions, belonging to \hat{l}^1 , and they have inverses that belong to \hat{l}^1 , also. The matrices \hat{M}_U and \hat{M}_V have the same dimensions as \hat{U} and \hat{V} , respectively, and they are of the form

$$(5) \quad \hat{M}_U = \begin{pmatrix} \hat{N}_U & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{M}_V = \begin{pmatrix} \hat{N}_V & 0 \\ 0 & 0 \end{pmatrix},$$

where the bordering zero matrices may be absent depending on the ranks of \hat{U} and \hat{V} , and \hat{N}_U and \hat{N}_V are diagonal matrices whose ranks at each point are the same as the ranks of the matrices \hat{U} and \hat{V} . In particular, by Assumption 1.1, the rank of $\hat{N}_U(z)$ and $\hat{N}_V(z)$ are constant on the unit circle $|z| = 1$. That such a decomposition exists for stable real rational matrices is well known, (see, e.g., [16, Thm. 29, p. 404]),

and in this case all the matrices will be stable real rational. It is also true in the case where \widehat{U} and \widehat{V} are analytic in a neighborhood of \overline{D} . However, it does not seem to be known whether the corresponding result for arbitrary ℓ^1 -valued matrices satisfying Assumption 1.1 is true. It is true if \widehat{U} and \widehat{V} have either full row rank or full column rank; this can be proved by modifying the argument in [8]. Therefore, in the case where \widehat{U} and \widehat{V} are ℓ^1 -valued we shall simply assume that the following assumption holds (we conjecture that this assumption is redundant).

Assumption 2.1. The matrices $\widehat{U}(z)$ and $\widehat{V}(z)$ have Smith factorizations of the form (4).

Let us proceed, taking Assumption 2.1 for granted. Since \widehat{R}_U and \widehat{L}_V are invertible ℓ^1 square matrices, they can be absorbed into \widehat{Q} (there is a one-to-one correspondence between \widehat{Q} and $\widehat{R}_U \widehat{Q} \widehat{L}_V$ in ℓ^1), and hence we may, without loss of generality, assume that \widehat{K} is of the form

$$(6) \quad \widehat{K} = \widehat{L}_U \widehat{M}_U \widehat{Q} \widehat{M}_V \widehat{R}_V.$$

If we partition \widehat{L}_U , \widehat{Q} , and \widehat{R}_V conformally with the partitions of \widehat{M}_U and \widehat{M}_V given above into

$$(7) \quad \widehat{L}_U = \begin{pmatrix} \widehat{L}_{U,1} & \widehat{L}_{U,2} \end{pmatrix}, \quad \widehat{Q} = \begin{pmatrix} \widehat{Q}_{11} & \widehat{Q}_{12} \\ \widehat{Q}_{21} & \widehat{Q}_{22} \end{pmatrix}, \quad \widehat{R}_V = \begin{pmatrix} \widehat{R}_{V,1} \\ \widehat{R}_{V,2} \end{pmatrix},$$

then we can write \widehat{K} in the form

$$(8) \quad \begin{aligned} \widehat{K} &= \begin{pmatrix} \widehat{L}_{U,1} & \widehat{L}_{U,2} \end{pmatrix} \begin{pmatrix} \widehat{N}_U & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \widehat{Q}_{11} & \widehat{Q}_{12} \\ \widehat{Q}_{21} & \widehat{Q}_{22} \end{pmatrix} \begin{pmatrix} \widehat{N}_V & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \widehat{R}_{V,1} \\ \widehat{R}_{V,2} \end{pmatrix} \\ &= \widehat{L}_{U,1} \widehat{N}_U \widehat{Q}_{1,1} \widehat{N}_V \widehat{R}_{V,1}. \end{aligned}$$

In this new representation the left multiplier $\widehat{L}_{U,1} \widehat{N}_U$ has full column rank and the right multiplier $\widehat{N}_V \widehat{R}_{V,1}$ has full row rank on the unit circle $|z| = 1$. Since they are analytic in the open unit disk D , it must be true that $\widehat{L}_{U,1} \widehat{N}_U$ has full column rank and $\widehat{N}_V \widehat{R}_{V,1}$ has full row rank in all but finitely many points in D . This proves our claim that we may almost, without loss of generality, assume that \widehat{U} has full column rank and that \widehat{V} has full row rank for all but finitely many $z \in \overline{D}$, and that none of the exceptional points lie on the unit circle $|z| = 1$. In particular, (5) becomes

$$(9) \quad \widehat{M}_U = \begin{pmatrix} \widehat{N}_U \\ 0 \end{pmatrix}, \quad \widehat{M}_V = \begin{pmatrix} \widehat{N}_V & 0 \end{pmatrix}.$$

As we shall see in (15), after the reduction of the dimension of Q described above has been carried out, Q is uniquely determined by K . This fact is important in the formulation of the model matching problem as an infinite-dimensional linear programming problem, and especially so in the multiblock case. In the general case we can solve Q , using the formulas above, as a unique function of K , $Q_{1,2}$, $Q_{2,1}$ and $Q_{2,2}$, where the last three are free parameters.

The reduction that we have described above is not always carried out in the one-block case; see, e.g., the discussion of the one-block case in [10].

3. Frequency domain formulation: Interpolation constraints. The model matching problem described can be formulated as a linear programming problem in

l^1 . The first step in this formulation is to construct a set of conditions that describe the class of all possible \mathcal{Z} -transforms \hat{K} in \hat{l}^1 that we get when Q varies over l^1 . Such descriptions are given in [4], [5], and [10]. However, since there are some redundancies in these descriptions, and since these redundancies have some nontrivial consequences for the main results of this paper, we prefer to give a slightly different formulation. Our formulation is an extension to the full multivariable case of the formulations in [14] and [15].

There is a set of necessary and sufficient conditions on a function $\hat{K} \in \hat{l}^1$ in order for it to be of the form $\hat{K} = \hat{U}\hat{Q}\hat{V}$ for some $\hat{Q} \in \hat{l}^1$. This set of conditions is simpler in the one-block case (the case where $n = p$ and $q = m$) than in the multiblock case. In the one-block case they consist of a finite number of “interpolation constraints,” and in the multiblock case we need both interpolation constraints and “convolution constraints.” We begin with a description of the former type of constraints. In the one-block case our description will be the same as one of the two alternative descriptions given in [10]. It is based on the Smith decompositions of \hat{U} and \hat{V} .

We begin our construction by partitioning the inverses of \hat{L}_U and \hat{R}_V into

$$(10) \quad \hat{L}_U^{-1} = \begin{pmatrix} \hat{U}_0 \\ \hat{U}_1 \end{pmatrix}, \quad \hat{R}_V^{-1} = \begin{pmatrix} \hat{V}_0 & \hat{V}_1 \end{pmatrix},$$

where the size of the different matrices correspond to the sizes of the matrices in (9), i.e., the dimensions of \hat{U}_0 and \hat{V}_0 are $p \times m$ and $n \times q$, respectively. Observe that the matrices \hat{U}_1 and \hat{V}_1 are absent in the one-block case, i.e., in this case $\hat{U}_0 = \hat{L}_U^{-1}$ and $\hat{V}_0 = \hat{R}_V^{-1}$. Multiply the identity $\hat{K} = \hat{L}_U \hat{M}_U \hat{R}_U \hat{Q} \hat{L}_V \hat{M}_V \hat{R}_V$ from the left by \hat{U}_0 and from the right by \hat{V}_0 to get (see (9))

$$(11) \quad \hat{U}_0 \hat{K} \hat{V}_0 = \hat{N}_U \hat{R}_U \hat{Q} \hat{L}_V \hat{N}_V.$$

This equation can be further analyzed as follows. Recall that \hat{N}_U and \hat{N}_V are diagonal matrices, i.e., $\hat{N}_U = \text{diag}(\hat{d}_{U,1}, \dots, \hat{d}_{U,m})$ and $\hat{N}_V = \text{diag}(\hat{d}_{V,1}, \dots, \hat{d}_{V,n})$. Partition \hat{U}_0 and \hat{V}_0 into sets of m row vectors and n column vectors, respectively, as

$$\hat{U}_0 = \begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_m \end{pmatrix} \quad \hat{V}_0 = \begin{pmatrix} \hat{\beta}_1 & \dots & \hat{\beta}_n \end{pmatrix}.$$

Then (11) becomes

$$(12) \quad \hat{\alpha}_i \hat{K} \hat{\beta}_j = \hat{d}_{U,i} \hat{q}_{i,j} \hat{d}_{V,j}, \quad i \in \{1, \dots, m\}, \quad j \in \{1, \dots, n\},$$

where $\hat{q}_{i,j}$ is the ij th element of $\hat{R}_U \hat{Q} \hat{L}_V$. Denote the set of all zeros of all different functions $\hat{d}_{U,i}$ and $\hat{d}_{V,j}$ in D by \mathcal{Z}_{UV} , and for each $z_0 \in \mathcal{Z}_{UV}$, let $\sigma_{U,i}(z_0)$ and $\sigma_{V,j}(z_0)$ denote the order of the zero of $\hat{d}_{U,i}$ and $\hat{d}_{V,j}$ at z_0 , respectively (we define $\sigma_{U,i}(z_0) = 0$ if $\hat{d}_{U,i}(z_0) \neq 0$, and $\sigma_{V,j}(z_0) = 0$ if $\hat{d}_{V,j}(z_0) \neq 0$).

From formula (12), together with the notations that we introduced, the following theorem follows more or less immediately (apart from the uniqueness and continuity claims, this is [10, Lem. 1]; these claims are not valid in [10, Lem. 1] due to the fact that there the reduction of §2 was not carried out).

THEOREM 3.1. *Given \hat{K} , \hat{U} , and \hat{V} in \hat{l}^1 of the type described above (i.e., we are in the one-block case, the dimensions are $m \times n$, $m \times m$, and $n \times n$, respectively,*

Assumption 1.1 is satisfied, and \widehat{U} and \widehat{V} have full rank in all but finitely many points in D), there is a unique function $\widehat{Q} \in \hat{l}^1$ such that $\widehat{K} = \widehat{U}\widehat{Q}\widehat{V}$ if and only if for all $z_0 \in \mathcal{Z}_{UV}$, all $i \in \{1, \dots, m\}$, and all $j \in \{1, \dots, n\}$

$$(13) \quad (\hat{\alpha}_i \widehat{K} \hat{\beta}_j)^{(r)}(z_0) = 0, \quad r \in \{0, \dots, \sigma_{U,i}(z_0) + \sigma_{V,j}(z_0) - 1\}.$$

Moreover, both the mapping $\widehat{Q} \mapsto \widehat{K} = \widehat{U}\widehat{Q}\widehat{V}$ and its inverse $\widehat{K} \mapsto \widehat{Q} = [\widehat{U}]^{-1}\widehat{K}[\widehat{V}]^{-1}$ are continuous.

Here (13) is considered to be vacuously satisfied if $\sigma_{U,i}(z_0) = \sigma_{V,j}(z_0) = 0$, i.e., if $\hat{d}_{U,i}(z_0) \neq 0$ and $\hat{d}_{V,j}(z_0) \neq 0$.

Proof. That (13) is a necessary condition is obvious from (12), and that it is sufficient follows from the discrete version of [7, Prop. 2.3].

Clearly, the operator that takes $Q \in l_{m \times n}^1$ into $U * Q * V \in l_{m \times n}^1$ is continuous. We have just characterized its range as the set of all functions in $l_{m \times n}^1$ that satisfy (13). This set of functions is closed in l^1 ; hence the range of the operator mentioned above is closed. By the open mapping theorem, the inverse of this operator, defined on its range, is continuous. \square

The total number σ_{UV} of conditions in (13) is

$$\sigma_{UV} = \sum_{z_0, i, j} (\sigma_{U,i}(z_0) + \sigma_{V,j}(z_0)) = \sigma_U n + \sigma_V m,$$

where σ_U and σ_V are the total multiplicities of the zeros of $\det \widehat{U}$ and $\det \widehat{V}$ in D , respectively, i.e.,

$$\sigma_U = \sum_{z_0 \in \mathcal{Z}_{UV}} \sum_{i \in \{1, \dots, m\}} \sigma_{U,i}(z_0), \quad \sigma_V = \sum_{z_0 \in \mathcal{Z}_{UV}} \sum_{j \in \{1, \dots, n\}} \sigma_{V,j}(z_0).$$

Furthermore, the conditions in (13) are linearly independent (one way to see this is to transfer the conditions back to the time domain, as is done in §5, and to observe that the resulting time domain conditions are linearly independent). Thus, the set of all possible elements of the form \widehat{K} is a closed subspace in \hat{l}^1 of finite codimension $\sigma_{UV} = \sigma_U n + \sigma_V m$.

If we replace \widehat{K} in (13) by $\widehat{H} - \widehat{\Phi}$, then we get

$$(14) \quad (\hat{\alpha}_i \widehat{\Phi} \hat{\beta}_j)^{(r)}(z_0) = (\hat{\alpha}_i \widehat{H} \hat{\beta}_j)^{(r)}(z_0), \quad r \in \{0, \dots, \sigma_{U,i}(z_0) + \sigma_{V,j}(z_0) - 1\}.$$

This equation can be interpreted as a requirement that $\widehat{\Phi}$ interpolates \widehat{H} in certain directions at each point $z_0 \in \mathcal{Z}_{UV}$. For this reason we shall refer to (13) and (14) as the *interpolation constraints*. In the H^∞ literature they are known as the two-sided Lagrange–Sylvester interpolation conditions ([1, §16.8]) or as the *discrete* interpolation conditions [2].

The definition of $\hat{\alpha}_i$ and $\hat{\beta}_j$ was based on the global Smith factorizations of \widehat{U} and \widehat{V} . For actual computations it is more convenient to use different functions $\hat{\alpha}_i$ and $\hat{\beta}_j$ at each different point $z_0 \in \mathcal{Z}_{UV}$, i.e., we use several *local* Smith factorizations, one at each point, rather than one global Smith factorization. This is possible because, at each z_0 , (14) depends only on the derivatives of $\hat{\alpha}_i$ and $\hat{\beta}_j$ of order at most $\sigma_{U,i}(z_0) + \sigma_{V,j}(z_0) - 1$. Thus, it is not necessary to use the same functions at all points $z_0 \in \mathcal{Z}_{UV}$, and we may always, without loss of generality, assume that at the point $z_0 \in \mathcal{Z}_{UV}$, the functions $\hat{\alpha}_i$ and $\hat{\beta}_j$ are constants if $\sigma_{U,i}(z_0) \leq 1$ and $\sigma_{V,j}(z_0) \leq 1$, respectively, and

that they are polynomials of order at most $\sigma_{U,i}(z_0) - 1$ and $\sigma_{V,j}(z_0) - 1$ otherwise. This simplifies the linear programming formulation of the problem to be discussed in §5.

Theorem 3.1 immediately gives us the following corollary.

COROLLARY 3.2. *In the one-block case, if \mathcal{Z}_{UV} is empty, or more generally, if for all $z_0 \in \mathcal{Z}_{UV}$, all $i \in \{1, \dots, m\}$, and all $j \in \{1, \dots, n\}$*

$$(\hat{\alpha}_i \hat{H} \hat{\beta}_j)^{(r)}(z_0) = 0, \quad r \in \{0, \dots, \sigma_{U,i}(z_0) + \sigma_{V,j}(z_0) - 1\}$$

then the optimal solution of (OPT) is $\Phi = 0$.

When this corollary is applied to the standard sensitivity minimization problem we recover the well-known fact that there are no restrictions on the achievable sensitivity for a minimum phase plant.

As proposed to us by one of the referees, let us end this section by pointing out a minor error in the discussion in [10] on the interpolation constraints. McDonald and Pearson present two different sets of interpolation constraints, one set in [10, Def. 2] and [10, Thm. 3], and another set in [10, Lem. 1]. The latter set is the one that we use in Theorem 3.1 and, as we have seen, it is linearly independent. This contradicts a claim made in [10], where it is stated that the first set of constraints is linearly independent, whereas the second set of constraints is not, and there it is recommended that we use the first set of constraints. To see that this claim must be false it suffices to compute the total number of scalar conditions contained in (13) on one hand, and in [10, Def. 2] on the other hand. As we have seen, the total number of scalar conditions in (13) is $\sigma_U n + \sigma_V m$. A similar computation shows that the number of scalar conditions in each of the parts (i), (ii), (iiia), and (iiib) of [10, Def. 2] are $\sigma_U n$, $\sigma_V m$, $\sigma_U n$, and $\sigma_V m$, respectively. When these numbers are added, we get either $2\sigma_U n + \sigma_V m$ or $\sigma_U n + 2\sigma_V m$, depending on which of the two alternative versions of (iii) that we use (if we use different versions of (iii) at different points, then we get some other number between these two). If \hat{U} and \hat{V} have no common zeros, then the two sets of constraints become equivalent (provided we use the trivially satisfied part of (iii); that is, at each point we use either (i) or (ii) and ignore all the remaining conditions), and it does not matter which set is used. However, the conditions listed [10, Def. 2] become redundant whenever \hat{U} and \hat{V} contain common zeros. One problem is that (i) and (ii) of [10, Def. 2] overlap in this case; to see this, note that they are equivalent to the requirement that for all $z_0 \in \mathcal{Z}_{UV}$, all $i \in \{1, \dots, m\}$, and all $j \in \{1, \dots, n\}$,

$$(\hat{\alpha}_i \hat{K})^{(r)}(z_0) \hat{\beta}_j(z_0) = 0, \quad r \in \{0, \dots, \sigma_{U,i}(z_0) - 1\}$$

and

$$\hat{\alpha}_i(z_0) (\hat{K} \hat{\beta}_j)^{(r)}(z_0) = 0, \quad r \in \{0, \dots, \sigma_{V,j}(z_0) - 1\},$$

respectively, and these conditions overlap already for $r = 0$ whenever both $\sigma_{U,i}(z_0)$ and $\sigma_{V,j}(z_0)$ are nonzero.

4. Frequency domain formulation: Convolution constraints. To see that the frequency domain conditions must be quite different in the multiblock case (i.e., two-block and four-block cases) compared to the single-block case, it suffices to look at a very simple example. Suppose that \hat{U} contains two identical rows. Then the corresponding rows of \hat{K} must also be identical. Clearly, the interpolation constraints

(13) that we described, and that completely characterize the one-block case, cannot force two rows of \hat{K} to be identical. Thus, we need some additional conditions for the multiblock case. In the previous example, there was a new constraint on \hat{K} of the following type: there is some vector $\alpha \in \mathbf{R}^m$ such that $\alpha\hat{K}$ vanishes identically. The general conditions that we shall use are generalizations of this simple condition, and they describe linear dependencies between the rows of \hat{U} and between the columns of \hat{V} .

Suppose that at least one of the two inequalities $p \leq m$ and $q \leq n$ is strict. Choose some arbitrary matrices \hat{U}_0 and \hat{V}_0 of dimension $p \times m$ and $n \times q$, respectively, such that $\hat{U}_0\hat{U}$ and $\hat{V}\hat{V}_0$ have rank p and q , respectively, in at least one point of D (if $p = n$ then we may choose $\hat{U}_0 = I$, and if $q = m$ then we may choose $\hat{V}_0 = I$). For example, one possible choice is to take U_0 and V_0 as in (10). This means that $\hat{U}_0\hat{U}$ and $\hat{V}\hat{V}_0$ are invertible in at least one point of D . Multiply the identity $\hat{K} = \hat{U}\hat{Q}\hat{V}$ from the left by $[\hat{U}_0\hat{U}]^{-1}\hat{U}_0$ and from the right by $\hat{V}_0[\hat{V}\hat{V}_0]^{-1}$ to get

$$(15) \quad \hat{Q} = [\hat{U}_0\hat{U}]^{-1}\hat{U}_0\hat{K}\hat{V}_0[\hat{V}\hat{V}_0]^{-1}.$$

The functions on the right-hand side are matrices with values in the quotient field of \hat{l}^1 (each element is a quotient of two elements of \hat{l}^1), and they determine \hat{Q} uniquely. Substituting this back into the equation $\hat{K} = \hat{U}\hat{Q}\hat{V}$ we get the following necessary condition that \hat{K} must satisfy:

$$(16) \quad \hat{K} = \hat{U}[\hat{U}_0\hat{U}]^{-1}\hat{U}_0\hat{K}\hat{V}_0[\hat{V}\hat{V}_0]^{-1}\hat{V}.$$

Formally, the matrix equation above contains mn different equations, one for each component of the matrix \hat{K} , but some of these equations are redundant. This redundancy is due to the fact that (16) has been constructed in such a way that the equation that we get by multiplying (16) from the left by \hat{U}_0 and from the right by \hat{V}_0 is always satisfied, independently of the choice of \hat{K} . We can remove this redundancy in the following way. Let \hat{U}_1 and \hat{V}_1 be matrices in \hat{l}^1 of dimensions $m - p$ and $n - q$, respectively, that complement \hat{U}_0 and \hat{V}_0 in the sense that the two matrices $\begin{pmatrix} \hat{U}_0 \\ \hat{U}_1 \end{pmatrix}$ and $\begin{pmatrix} \hat{V}_0 \\ \hat{V}_1 \end{pmatrix}$ are invertible in at least one point of D (one possible choice is the one in (10)). Then (16) is equivalent to the equation that we get when we multiply (16) from the left by $\begin{pmatrix} \hat{U}_0 \\ \hat{U}_1 \end{pmatrix}$ and from the right by $\begin{pmatrix} \hat{V}_0 \\ \hat{V}_1 \end{pmatrix}$. When this is done, we get the equivalent system

$$(17) \quad \begin{pmatrix} \hat{U}_0\hat{K}\hat{V}_0 & \hat{U}_0\hat{K}\hat{V}_1 \\ \hat{U}_1\hat{K}\hat{V}_0 & \hat{U}_1\hat{K}\hat{V}_1 \end{pmatrix} - \begin{pmatrix} \hat{U}_0\hat{K}\hat{V}_0 & (\hat{U}_0\hat{K}\hat{V}_0)[\hat{V}\hat{V}_0]^{-1}\hat{V}\hat{V}_1 \\ \hat{U}_1\hat{U}[\hat{U}_0\hat{U}]^{-1}(\hat{U}_0\hat{K}\hat{V}_0) & \hat{U}_1\hat{U}[\hat{U}_0\hat{U}]^{-1}(\hat{U}_0\hat{K}\hat{V}_0)[\hat{V}\hat{V}_0]^{-1}\hat{V}\hat{V}_1 \end{pmatrix} = 0,$$

which expresses $\hat{U}_0\hat{K}\hat{V}_1$, $\hat{U}_1\hat{K}\hat{V}_0$, and $\hat{U}_1\hat{K}\hat{V}_1$ as functions of $\hat{U}_0\hat{K}\hat{V}_0$. From this equation it is evident that the number of independent equations is $mn - pq$. (In the one-block case $mn - pq = 0$, and no extra conditions are needed).

We can turn (17) into an equation where all the factors belong to \hat{l}^1 (instead of to the quotient field of \hat{l}^1) by multiplying the equations by, e.g., the determinants of $\hat{U}_0\hat{U}$ and $\hat{V}\hat{V}_0$. Usually a better choice is to choose some other $p \times p$ and $q \times q$ -dimensional matrix functions \hat{W}_0 and \hat{W}_1 , respectively, with full rank at almost all points of D ,

in such a way that $\widehat{W}_0 \widehat{U}_1 \widehat{U} [\widehat{U}_0 \widehat{U}]^{-1}$ and $[\widehat{V} \widehat{V}_0]^{-1} \widehat{V} \widehat{V}_1 \widehat{W}_1$ have no singularities, and to multiply (17) from the left by $\begin{pmatrix} I & 0 \\ 0 & \widehat{W}_0 \end{pmatrix}$ and from the right by $\begin{pmatrix} I & 0 \\ 0 & \widehat{W}_1 \end{pmatrix}$. However, this leads to a system of the same type as (17); the only difference is that \widehat{U}_1 has been replaced by $\widehat{W}_0 \widehat{U}_1$ and that \widehat{V}_1 has been replaced by $\widehat{V}_1 \widehat{W}_1$. Thus, this discussion shows that it is possible to choose \widehat{U}_1 and \widehat{V}_1 in (17) in such a way that all the factors multiplying \widehat{K} belong to \hat{l}^1 . Throughout in the sequel we assume that this has been done. (For further comments on the choice of multipliers, see the discussion at the end of this section, as well as Lemma 6.4.)

If we transform (17) back into the time domain, then we get $mn - pq$ linearly independent convolution equations. For this reason we shall refer to the equations (16) and (17) as the *convolution constraints*. (In the H^∞ -literature these are called *continuous* interpolation conditions; see [2].) In the linear programming approach that is presented later we employ (17), but for simplicity we shall base the rest of the discussion in this section on (16).

An important observation is that the convolution constraints (16) are independent of \widehat{U}_0 and \widehat{V}_0 in the sense that if they are satisfied for one pair of functions \widehat{U}_0 and \widehat{V}_0 , then they are satisfied for every other pair of functions \widehat{U}_0 and \widehat{V}_0 , also, as long as $\widehat{U}_0 \widehat{U}$ and $\widehat{V} \widehat{V}_0$ are invertible in at least one point of D . To see this, multiply (16) from the left by $[\widehat{U}_0 \widehat{U}]^{-1} \widehat{U}_0$ and from the right by $\widehat{V}_0 [\widehat{V} \widehat{V}_0]^{-1}$ to get

$$(18) \quad [\widehat{U}_0 \widehat{U}]^{-1} \widehat{U}_0 \widehat{K} \widehat{V}_0 [\widehat{V} \widehat{V}_0]^{-1} = [\widehat{U}_0 \widehat{U}]^{-1} \widehat{U}_0 \widehat{K} \widehat{V}_0 [\widehat{V} \widehat{V}_0]^{-1}.$$

This implies two things. First, (16) is valid with \widehat{U}_0 and \widehat{V}_0 replaced by \widehat{U}_0 and \widehat{V}_0 , and second, if we use (15) as a definition of \widehat{Q} , then \widehat{Q} is independent of the particular choice of \widehat{U}_0 and \widehat{V}_0 .

Let us collect this argument into the following lemma.

LEMMA 4.1. *Let \widehat{U} , and \widehat{V} be matrices of dimension $m \times p$, and $q \times n$, respectively, with values in \hat{l}^1 , and suppose that \widehat{U} has full column rank and that \widehat{V} has full row rank.*

(i) *If \widehat{K} is a matrix of the form $\widehat{K} = \widehat{U} \widehat{Q} \widehat{V}$ for some $p \times q$ -dimensional matrix \widehat{Q} with values in \hat{l}^1 , then \widehat{K} satisfies the convolution constraints (16) for all matrices \widehat{U}_0 and \widehat{V}_0 of dimensions $p \times m$ and $n \times q$, respectively, with values in \hat{l}^1 , such that $\widehat{U}_0 \widehat{U}$ and $\widehat{V} \widehat{V}_0$ are invertible in at least one point of D .*

(ii) *Conversely, suppose that \widehat{K} is an $m \times n$ -dimensional matrix with values in \hat{l}^1 that satisfies the convolution constraints (16) for one pair of functions \widehat{U}_0 and \widehat{V}_0 of the type mentioned in (i). Then it satisfies the same constraints for all other possible choices of \widehat{U}_0 and \widehat{V}_0 , also, and it determines a function \widehat{Q} uniquely through the equations (15), where the choice of \widehat{U}_0 and \widehat{V}_0 is irrelevant in the sense that \widehat{Q} is independent of the particular choice of \widehat{U}_0 and \widehat{V}_0 . Each element of the function \widehat{Q} is the quotient of two functions belonging to \hat{l}^1 .*

As we mentioned previously, it is possible to choose U_0 , U_1 , V_0 , and V_1 as in (10). In this case

$$(19) \quad [\widehat{U}_0 \widehat{U}]^{-1} = [\widehat{R}_U]^{-1} [\widehat{N}_U]^{-1}, \quad [\widehat{V} \widehat{V}_0]^{-1} = [\widehat{N}_V]^{-1} [\widehat{L}_V]^{-1},$$

and (16) becomes (cf. (7))

$$(20) \quad \widehat{K} = \widehat{L}_{U,1} \widehat{U}_0 \widehat{K} \widehat{V}_0 \widehat{R}_{V,1}.$$

From this version of (16) it is easy to prove the following result.

LEMMA 4.2. *A function $\hat{K} \in \hat{\mathbb{I}}_{m \times n}^1$ satisfies (16) if and only if it is of the form $\hat{K} = \hat{L}_{U,1} \hat{S} \hat{R}_{V,1}$ for some $\hat{S} \in \hat{\mathbb{I}}_{p \times q}^1$. Moreover, both the mapping $\hat{S} \mapsto \hat{K} = \hat{L}_{U,1} \hat{S} \hat{R}_{V,1}$ and its inverse $\hat{K} \mapsto \hat{S} = \hat{U}_0 \hat{K} \hat{V}_0$, where \hat{U}_0 and \hat{V}_0 satisfy (10), is continuous.*

Proof. By (20), the condition $\hat{K} = \hat{L}_{U,1} \hat{S} \hat{R}_{V,1}$ is necessary (choose \hat{S} to be $\hat{S} = \hat{U}_0 \hat{K} \hat{V}_0$). Conversely, if \hat{S} is of this type, then we can multiply the identity $\hat{K} = \hat{L}_{U,1} \hat{S} \hat{R}_{V,1}$ from the left by \hat{U}_0 and from the right by \hat{V}_0 to get $\hat{S} = \hat{U}_0 \hat{K} \hat{V}_0$. This, substituted back into the equation $\hat{K} = \hat{L}_{U,1} \hat{S} \hat{R}_{V,1}$ shows that \hat{K} satisfies (20). The final continuity claim is obvious. \square

Let us return to the interpolation constraints that we found in §3. For later use, let us record the fact that the set of all the solutions of (13) can be parametrized in a way similar to the parametrization presented in Lemma 4.2.

LEMMA 4.3. *A function $\hat{K} \in \hat{\mathbb{I}}_{m \times n}^1$ satisfies (13) if and only if it is of the form*

$$(21) \quad \hat{K} = (\hat{L}_{U,1} \quad \hat{L}_{U,2}) \begin{pmatrix} \hat{N}_U \hat{T}_{11} \hat{N}_V & \hat{T}_{12} \\ \hat{T}_{21} & \hat{T}_{22} \end{pmatrix} \begin{pmatrix} \hat{R}_{V,1} \\ \hat{R}_{V,2} \end{pmatrix}$$

for some matrices $\hat{T}_{ij} \in \hat{\mathbb{I}}^1$ of appropriate dimensions. The mapping that takes $\begin{pmatrix} \hat{T}_{11} & \hat{T}_{12} \\ \hat{T}_{21} & \hat{T}_{22} \end{pmatrix}$ into K is continuous.

Proof. Let K satisfy (21). Multiply (21) from the left by \hat{U}_0 and from the right by \hat{V}_0 , where \hat{U}_0 and \hat{V}_0 satisfy (10), to get

$$\hat{U}_0 \hat{K} \hat{V}_0 = \hat{N}_U \hat{T}_{11} \hat{N}_V.$$

Clearly, this implies that K satisfies (13).

Conversely, if K satisfies (13), then the function \hat{T}_{11} defined by

$$\hat{T}_{11} = [\hat{N}_U]^{-1} \hat{U}_0 \hat{K} \hat{V}_0 [\hat{N}_V]^{-1}$$

belongs to $\hat{\mathbb{I}}^1$. Define

$$\hat{T}_{12} = \hat{U}_0 \hat{K} \hat{V}_1, \quad \hat{T}_{21} = \hat{U}_1 \hat{K} \hat{V}_0, \quad \hat{T}_{22} = \hat{U}_1 \hat{K} \hat{V}_1,$$

to get the desired representation. \square

By combining the preceding lemmas we get the following theorem.

THEOREM 4.4. *Given \hat{K} , \hat{U} , and \hat{V} in $\hat{\mathbb{I}}^1$ of the type described above (i.e., we are in the multiblock case, the dimensions are $m \times n$, $m \times p$, and $q \times n$, respectively, Assumption 1.1 is satisfied, and \hat{U} and \hat{V} have full column rank and row rank, respectively, in all but finitely many points in D), there is a unique function $\hat{Q} \in \hat{\mathbb{I}}^1$ such that $\hat{K} = \hat{U} \hat{Q} \hat{V}$ if and only if \hat{K} satisfies the convolution constraints (16) as explained in Lemma 4.1 and, in addition, for all $z_0 \in Z_{UV}$, all $i \in \{1, \dots, m\}$, and all $j \in \{1, \dots, n\}$, (13) holds. Moreover, both the mapping $\hat{Q} \mapsto \hat{K} = \hat{U} \hat{Q} \hat{V}$ and its inverse (defined on the range) are continuous. The inverse mapping has the representation (15). In particular, if \hat{U}_0 and \hat{V}_0 are chosen according to the decomposition (10), then this inverse mapping can be written in the form*

$$\hat{K} \mapsto \hat{Q} = [\hat{R}_U]^{-1} [\hat{N}_U]^{-1} \hat{U}_0 \hat{K} \hat{V}_0 [\hat{N}_V]^{-1} [\hat{L}_V]^{-1}.$$

The proof of this theorem is similar to the earlier proofs, and we leave it to the reader.

The comments at the end of §3 on the freedom of choice of the vectors $\hat{\alpha}_i$ and $\hat{\beta}_j$ also apply here. In particular, at each point $z \in \mathcal{Z}_{UV}$ we only need to know the particular local zero structure at that point, and α_i and β_j may be chosen to be polynomials, different at each point $z \in \mathcal{Z}_{UV}$. There is also a large freedom in the choice of the particular \hat{U}_0 and \hat{V}_0 to be used in Lemma 4.1. One choice that seems to be particularly attractive is to take \hat{U}_0 , \hat{U}_1 , \hat{V}_0 , and \hat{V}_1 to be determined from the Smith factorization of \hat{U} and \hat{V} , as in (10). Then

$$\hat{U}_0 \hat{U} = \hat{N}_U \hat{R}_U, \quad \hat{U}_1 \hat{U} = 0, \quad \hat{V} \hat{V}_0 = \hat{L}_V \hat{N}_V, \quad \hat{V} \hat{V}_1 = 0,$$

and (17) simplifies into

$$(22) \quad \begin{pmatrix} 0 & \hat{U}_0 \hat{K} \hat{V}_1 \\ \hat{U}_1 \hat{K} \hat{V}_0 & \hat{U}_1 \hat{K} \hat{V}_1 \end{pmatrix} = 0.$$

Another somewhat more flexible approach is to choose \hat{U}_0 and \hat{V}_0 in such a way that the factors $\hat{U}[\hat{U}_0 \hat{U}]^{-1}$ and $[\hat{V} \hat{V}_0]^{-1} \hat{V}$ appearing in (17) belong to \hat{l}^1 . Observe that we can rewrite these factors, using the Smith factorizations of \hat{U} and \hat{V} , into the form

$$(23) \quad \hat{U}[\hat{U}_0 \hat{U}]^{-1} = \hat{L}_{U,1}[\hat{U}_0 \hat{L}_{U,1}]^{-1}, \quad [\hat{V} \hat{V}_0]^{-1} \hat{V} = [\hat{R}_{V,1} \hat{V}_0]^{-1} \hat{R}_{V,1}.$$

If (10) holds, then $\hat{U}[\hat{U}_0 \hat{U}]^{-1} = \hat{L}_{U,1}$ and $[\hat{V} \hat{V}_0]^{-1} \hat{V} = \hat{R}_{V,1}$. However, it really suffices to choose \hat{U}_0 and \hat{V}_0 in such a way that $\hat{U}_0 \hat{L}_{U,1}$ and $\hat{R}_{V,1} \hat{V}_0$ are invertible. If this is done, then it is possible to choose the matrices \hat{U}_1 and \hat{V}_1 in such a way that $\begin{pmatrix} \hat{U}_0 \\ \hat{U}_1 \end{pmatrix}$ and $\begin{pmatrix} \hat{V}_0 & \hat{V}_1 \end{pmatrix}$ are invertible in all of \overline{D} . This condition becomes important in §11.

The theorem above should be compared to the corresponding Theorem 6 in [10]. The conditions (i) and (ii) used in [10, Thm. 6] are a special case of our equation (16). To get those conditions we choose the pair \hat{U}_0 and \hat{V}_0 in a special way, namely, they are chosen to be matrices that pick out p linearly independent rows from \hat{U} and q linearly independent columns from \hat{V} , respectively. The same choice of \hat{U}_0 and \hat{V}_0 are present in McDonald and Pearson's definition of the interpolation constraints for the multiblock case, also: instead of using the Smith factorizations of the full matrices \hat{U} and \hat{V} as we do they use the Smith factorizations of $\hat{U}_0 \hat{U}$ and $\hat{V} \hat{V}_0$ to define the vector functions $\hat{\alpha}_i$ and $\hat{\beta}_j$ that appear in (13). This leads to a result that differs significantly from ours, even if we ignore the redundancy in [10] due to the use of an alternative set of interpolation constraints (see the discussion at the end of the previous section). For one thing, the approach in [10] necessitates one extra nonstandard assumption, namely that there are p rows of \hat{U} and q columns of \hat{V} that are independent for all z on the unit circle (see Assumption 3 in [10]). This assumption is stronger than our Assumption 1.1. Moreover, since those specifically chosen rows of \hat{U} and columns of \hat{V} may lose rank at additional points in D , not belonging to the set \mathcal{Z}_{UV} determined by the zero structure of the full matrices \hat{U} and \hat{V} , in McDonald and Pearson's formulation there are usually additional points $z \in D$ that do not belong to \mathcal{Z}_{UV} , but that have to be included among the points z_0 in (13). It may also be the case that, at a particular point $z \in \mathcal{Z}_{UV}$, the determinants of the chosen rows and columns vanish to a higher order than the true total multiplicity of that point, and this leads to a greater number of derivatives in (13) than is needed in our formulation. In other words, the conditions listed in [10, Thm. 6] are, not in general, linearly independent. As reported in [14] and [15], such linear dependencies

may cause numerical instabilities in the solution schemes and, as we shall see in §11, it leads to nonuniqueness of the dual solutions. Moreover, with this choice of \widehat{U}_0 and \widehat{V}_0 it is not in general possible to choose \widehat{U}_1 and \widehat{V}_1 in such a way that $\begin{pmatrix} \widehat{U}_0 \\ \widehat{U}_1 \end{pmatrix}$ and $\begin{pmatrix} \widehat{V}_0 \\ \widehat{V}_1 \end{pmatrix}$ are invertible in all of \overline{D} ; this property seems to be highly desirable (see §11).

There is one case in particular where the difference between the method presented here and the one in [10] becomes profound. For a certain type of problems Z_{UV} is empty, whereas the corresponding set in McDonald and Pearson's formulation may be nonempty; see [13]. This has a great influence on the resulting linear programming system that we shall introduce in §6. Moreover, in this case, with our formulation the resulting FME method (see §§8 and 9) becomes very easy to implement since the critical computation of the value of N drops out.

5. Linear programming formulation: One-block case. The frequency domain descriptions that we have given can be transformed back into the time domain. The time domain formulation contains infinitely many variables, i.e., the elements $\Phi_{ij}(k)$ of the matrix sequence $\Phi = H - K = H - U * Q * V$. In the one-block case it contains finitely many equations, more precisely, σ_{UV} equations, and in the multiblock case it contains infinitely many equations. By now this transformation is standard, and it has been used in, e.g., [4], [5], [10], [11], [14], and [15]. However, none of these references contains all the different aspects of the time domain formulation that we shall need here. The paper [4] contains a nice description of the one-block case, including both the primal and the dual problems, but there only simple zeros are allowed, and the multiblock case is missing. In the multiblock papers the descriptions of the primal problems are satisfactory, but some of the aspects of the formulation of the dual problem are missing. Therefore, for the convenience of the reader we have included a short description of this transformation. For more details, especially for the primal formulation, we refer the reader to the references listed. In this section we give a description of the one-block case (similar to the description in [4]), and defer the discussion of the two-block case to the next section.

First we transform the interpolation constraints (14) back into the time domain and get (to simplify the notations we have replaced α_i by α and β_j by β , and use i and j as summation indices)

$$\sum_{j=0}^{\infty} j^r z_0^j \sum_{i=0}^j \alpha(j-i) \sum_{k=0}^i (\Phi(k) - H(k)) \beta(i-k) = 0$$

for all relevant α , β , z_0 , and r . After a change of the order of summation this becomes

$$(24) \quad \sum_{k=0}^{\infty} \sum_{i=k}^{\infty} \sum_{j=i}^{\infty} j^r z_0^j \alpha(j-i) (\Phi(k) - H(k)) \beta(i-k) = 0.$$

The left-hand side of this equation is a continuous affine mapping from the set of $m \times n$ -dimensional sequences Φ in l^1 into \mathbb{C} . Let us denote the coefficient of $\Phi(k)$ by $A(k)$, where $A(k)$ is a second-order tensor of dimensions $m \times n$. Then (24) becomes

$$(25) \quad \sum_{k=0}^{\infty} A(k) \Phi(k) = \sum_{k=0}^{\infty} A(k) H(k),$$

where the products $A(k)\Phi(k)$ and $A(k)H(k)$ are the usual tensor products

$$A(k)\Phi(k) = \sum_{ij} A_{ij}(k) \Phi_{ij}(k), \quad A(k)H(k) = \sum_{ij} A_{ij}(k) H_{ij}(k).$$

The coefficients $A(k)$ have one crucial property, namely, they satisfy $|A(k)| \rightarrow 0$ as $k \rightarrow \infty$; this follows directly from (24) and the fact that $|z_0| < 1$.

Above we discussed the inversion of (14) for one particular choice of α , β , z_0 , and r . The complete set of conditions (14) contains σ_{UV} different conditions. When they all are inverted, we can write the result in the form

$$(26) \quad \sum_{k=0}^{\infty} A(k) \Phi(k) = a,$$

where each $A(k)$ is a third-order tensor, mapping the space of $m \times n$ -dimensional matrices into a complex σ_{UV} -dimensional vector, and $a = \sum_{k=0}^{\infty} A(k) H(k)$ is an σ_{UV} -dimensional vector with complex entries. Observe that, since we start out with real-valued data, if $z_0 \in \mathcal{Z}_{UV}$, then $\bar{z}_0 \in \mathcal{Z}_{UV}$. This means that if the set of equations above is not real, then the nonreal equations appear in complex conjugate pairs. By replacing each such pair by the real part and the imaginary part of one of the two equations we may assume that the entries of A and a are real.

Let us record the preceding argument into the following lemma.

LEMMA 5.1. *The interpolation constraints (14) are equivalent to an equation of the form (26), where each $A(k)$ is a third-order tensor of dimension $\sigma_{UV} \times (m \times n)$ satisfying $|A(k)| \rightarrow 0$ as $k \rightarrow \infty$, $a = \sum_{k=0}^{\infty} A(k) H(k)$ is a σ_{UV} -dimensional vector with real entries, and the products of $A(k)$ with $\Phi(k)$ and $H(k)$ are interpreted as $(A(k) \Phi(k))_{\ell} = \sum_{ij} A_{\ell ij}(k) \Phi_{ij}(k)$ and $(A(k) H(k))_{\ell} = \sum_{ij} A_{\ell ij}(k) H_{ij}(k)$.*

As a byproduct of this result we get the following corollary (the same result could have been deduced directly from Theorem 3.1, with \hat{K} replaced by $\hat{H} - \hat{\Phi}$).

COROLLARY 5.2. *Under Assumption 1.1, in the one-block case we may, without loss of generality, assume that \hat{H} , \hat{U} , and \hat{V} are polynomials in the sense that the original problem with data in \hat{l}^1 is equivalent to a problem with polynomial data.*

To prove this it suffices to observe that we get the same equations (26) if we, instead of using the original \hat{H} , \hat{U} , and \hat{V} in the formulation of (26), use polynomial functions \hat{U} and \hat{V} that have the same left and right zero structure, respectively, as the original functions \hat{U} and \hat{V} have, and use a polynomial \hat{H} that interpolates the original \hat{H} to a sufficiently high degree at each point of \mathcal{Z}_{UV} .

To complete the formulation of (OPT) in the one-block case as a linear programming problem in l^1 we must supplement (26) with a set of inequalities describing the norm (2) that is suppose to be minimized. Clearly, this minimization is equivalent to the minimization of μ under the m constraints $\mu \geq \sum_{j,k} |\Phi_{ij}(k)|$, $i \in \{1, \dots, m\}$. Split each Φ into $\Phi^+ - \Phi^-$, where all the elements of Φ^+ and Φ^- are nonnegative. (A natural additional requirement would be that for each i , j , and k , either $\Phi_{ij}^+(k) = 0$ or $\Phi_{ij}^-(k) = 0$, but that requirement is difficult to incorporate in the formulation due to its nonlinearity.) Then the constraints related to the new cost variable μ can be written in the form

$$\mathbf{1}\mu \geq E(\Phi^+ + \Phi^-),$$

where $\mathbf{1}$ represents an m -dimensional column vector with all elements equal to one, and E is the tensor mapping Φ into a vector whose i th element is $\sum_{j,k} \Phi_{ij}(k)$. Observe that the coefficients of $\Phi^+(k)$ and $\Phi^-(k)$ in this sum are independent of k ; in particular they do not tend to zero as $k \rightarrow \infty$, as opposed to those in (26).

Collecting the different pieces, we get the following formulation of (OPT) in the one-block case:

$$\begin{aligned}
 (\text{PRIM}_1) \quad & \text{minimize } \mu, \\
 & \mathbf{1}\mu - E(\Phi^+ + \Phi^-) \geq 0, \\
 & A(\Phi^+ - \Phi^-) = a, \\
 & \Phi^+, \Phi^- \geq 0,
 \end{aligned}$$

where the unknown Φ^+ and Φ^- are $m \times n$ -dimensional matrices in l^1 , E is an operator from this space into \mathbf{R}^m , A is an operator from the same space into $\mathbf{R}^{\sigma_{UV}}$, and $a = AH \in \mathbf{R}^{\sigma_{UV}}$.

We claim that the dual of the preceding problem is another linear programming problem which is posed in l^∞ . We can get this dual problem from the problem above by proceeding formally in the same way as we would do in the finite-dimensional case, but is easier to justify the derivation of the dual problem if we proceed in a different way, i.e., in the same way as in [4]. Observe that the problem that we have posed above is equivalent to finding the distance of the sequence $H \in l^1$ to $\mathcal{N}(A)$; the null space of A (recall that $\Phi = H - K$, where H is fixed and $K \in \mathcal{N}(A)$). The standard dual problem is then the following (see, e.g., [9, Thm. 1, p. 119]):

$$\begin{aligned}
 & \text{maximize } \nu = \langle H, \delta \rangle, \\
 & \|\delta\|_{l^\infty} \leq 1, \\
 & \delta \in \mathcal{N}(A)^\perp.
 \end{aligned}$$

Here

$$\|\delta\|_{l^\infty} \stackrel{\text{def}}{=} \sum_{i=1}^m \left(\sup_{1 \leq j \leq n, k \geq 0} |\delta_{ij}(k)| \right) \leq 1$$

is the norm in l^∞ that is dual to the norm $\|\cdot\|_{l^1}$ defined earlier in (2).

To write the norm constraint on δ into linear programming form we may introduce new variables τ_i , $i \in \{1, \dots, m\}$, and require that

$$(27) \quad -\tau_i \leq \delta_{ij}(k) \leq \tau_i, \quad i = 1, \dots, m, \quad j = 1, \dots, n, \quad k = 0, 1, 2, \dots,$$

and that

$$(28) \quad \sum_{i=1}^m \tau_i = 1.$$

If we further observe that $\sum_{i=1}^m \tau_i = \mathbf{1}^* \tau$, and that $(E^* \tau)_{ij} = \tau_i$, then we arrive at an intermediate form

$$\begin{aligned}
 (\text{DUAL}'_1) \quad & \text{maximize } \nu = \langle H, \delta \rangle, \\
 & \mathbf{1}^* \tau = 1, \\
 & -E^* \tau \leq \delta \leq E^* \tau, \\
 & \delta \in \mathcal{N}(A)^\perp
 \end{aligned}$$

of the dual problem.

The preceding form is not yet the one that we would like to have. However, before we can proceed any further we must prove the following claim.

LEMMA 5.3. *Define A as above, and let $A^* : \mathbf{R}^{\sigma_{UV}} \rightarrow l_{m \times n}^\infty$ be the adjoint of A . Then $\mathcal{R}(A^*) \subset c_{m \times n}^0$ is finite-dimensional, hence closed in c^0 and weak* closed in l^∞ , A is the adjoint of A^* , and $\mathcal{N}(A)$ is weak*-closed in l^1 . In particular, $\mathcal{R}(A^*) = \mathcal{N}(A)^\perp = {}^\perp \mathcal{N}(A)$, and $\mathcal{N}(A) = {}^\perp \mathcal{R}(A^*) = \mathcal{R}(A^*)^\perp$.*

Parts of this lemma are found implicitly in some of the proofs given in [4], [5], and [10]. To understand the statement of Lemma 5.3 is important to keep in mind that l^∞ is the dual of l^1 which is the dual of c^0 . In particular, $\mathcal{N}(A)^\perp$ and ${}^\perp \mathcal{R}(A^*)$ are interpreted within the duality of l^1 and l^∞ , whereas ${}^\perp \mathcal{N}(A)$ and $\mathcal{R}(A^*)^\perp$ are interpreted within the duality of c^0 and l^1 .

Proof. Each element in the range of A^* is of the form $H_{ij}(k) = \sum_{r=1}^{\sigma_{UV}} \eta_r A_{rij}(k)$ for some $\eta \in \mathbf{R}^{\sigma_{UV}}$. Recall that for all r, i , and j , $|A_{rij}(k)| \rightarrow 0$ as $k \rightarrow \infty$. This implies that $\mathcal{R}(A^*) \subset c^0$, as claimed. Thus, A is the adjoint of A^* . Moreover, $\mathcal{R}(A^*)$ is finite-dimensional; hence it is closed in c^0 and weak*-closed in l^∞ . This implies that $\mathcal{R}(A^*) = \mathcal{N}(A)^\perp = {}^\perp \mathcal{N}(A)$, and that $\mathcal{N}(A) = {}^\perp \mathcal{R}(A^*) = \mathcal{R}(A^*)^\perp$. \square

Since, according to Lemma 5.3, $\mathcal{N}(A)^\perp = \mathcal{R}(A^*)$, we may replace $\delta \in \mathcal{N}(A)^\perp$ in (DUAL₁) by $A^*\eta$, and let η range over $\mathbf{R}^{\sigma_{UV}}$. This leads to our final version of the dual problem (recall that we have denoted AH by a):

$$\begin{aligned} \text{maximize } \nu &= \langle a, \eta \rangle, \\ \text{(DUAL}_1\text{)} \quad \mathbf{1}^* \tau &= 1, \\ -E^* \tau &\leq A^* \eta \leq E^* \tau. \end{aligned}$$

Observe that this problem is again in linear programming form. We urge every reader familiar with duality in finite-dimensional linear programming to check that, in a finite-dimensional setting, this is exactly the problem that we would get from (PRIM₁) by formally converting it to the dual problem according to the standard duality rules of linear programming.

We have posed the dual problems (DUAL₁') and (DUAL₁) in l^∞ , and considered them to be the dual of the problem (PRIM₁), posed in l^1 . If we instead posed (DUAL₁') and (DUAL₁) in c^0 , and compute their dual problem in l^1 , then we get back the problem (PRIM₁); to see this, revert the preceding computations and use Lemma 5.3. Since dual distance problems always have optimal solutions (see [9, Thm. 1, p. 119 and Thm. 2, p. 121]), we have the following result of the existence of solutions in the one-block case.

THEOREM 5.4. *In the situation described in Theorem 3.1, (the one-block case) problem (OPT) is equivalent to the infinite-dimensional linear programming problem (PRIM₁), and it has a minimizer $\Phi \in l^1$. The minimum is equal to the distance in l^1 of H to the weak*-closed null space of A , which is of finite codimension σ_{UV} . To compute this distance we may instead solve the dual problem (DUAL₁), which has a maximizer $\eta \in \mathbf{R}^{\sigma_{UV}}$.*

This is essentially the same result as [4, Thm. 4] and [10, Thm. 4]. The main difference of Theorem 5.4 compared to [10, Thm. 4] is that, since we use a linearly independent set of conditions in (13), we are able to identify the precise codimension of the null space of A , i.e., we are able to use a dual variable η of smallest possible dimension σ_{UV} . (See the discussion at the end of §3.)

6. Linear programming formulation: Multiblock case. In the multiblock case we have the same equations that we had above, but in addition we have an

infinite number of equations that we get by inverting (17). This leads to an additional equation of the form (again replace \widehat{K} by $\widehat{H} - \widehat{\Phi}$)

$$B * \Phi = b,$$

where $(B * \Phi)(\ell) = \sum_{k=0}^{\ell} B(\ell - k)\Phi(k)$, $b(\ell) = (B * H)(\ell) = \sum_{k=0}^{\ell} B(\ell - k)H(k)$, and each $B(\ell)$ is a third-order tensor mapping the space of $m \times n$ -dimensional matrices into \mathbf{R}^{mn-pq} (for a more detailed description of this operator, see the proof of Lemma 6.4). The important fact about B is that $B \in l^1$. When this equation is added to (PRIM₁), we get the linear programming formulation for the multiblock case:

$$\begin{aligned} & \text{minimize } \mu, \\ & \mathbf{1}\mu - E(\Phi^+ + \Phi^-) \geq 0, \\ (\text{PRIM}) \quad & A(\Phi^+ - \Phi^-) = a, \\ & B * (\Phi^+ - \Phi^-) = b, \\ & \Phi^+, \Phi^- \geq 0, \end{aligned}$$

where the meanings of Φ^+ , Φ^- , E , A , and a are the same as before, the operator $B*$ is a convolution operator mapping $m \times n$ -dimensional l^1 into $(mn - pq)$ -dimensional l^1 , and $b = B * H$ belongs to the range of $B*$.

We can make a computation similar to the one that we made in the one-block case to get a dual linear programming problem. Denote the operator that maps Φ into the pair $[A\Phi, B * \Phi]$ by $[A, B*]$. Then the equations in (PRIM) involving A and B say that $\Phi - H \in \mathcal{N}([A, B*])$. Thus, the standard dual problem is

$$\begin{aligned} & \text{maximize } \nu = \langle H, \delta \rangle, \\ & \mathbf{1}^* \tau = 1, \\ (\text{DUAL}') \quad & -E^* \tau \leq \delta \leq E^* \tau, \\ & \delta \in \mathcal{N}([A, B*])^\perp \end{aligned}$$

where we have replaced the condition $\|\delta\|_{l^\infty} \leq 1$ by the same inequalities as in (DUAL'₁). The adjoint $[A, B*]^*$ of the operator $[A, B*]$ is the operator that maps $(\eta, \gamma) \in \mathbf{R}^{\sigma_{UV}} \times l_{(mn-pq)}^\infty$ into $A^* \eta + B^* * \gamma \in l_{m \times n}^\infty$, where the adjoint $B^* *$ of $B*$ is the operator

$$(29) \quad (B^* * \gamma)(k) = \sum_{\ell=k}^{\infty} B^*(\ell - k)\gamma(\ell).$$

Thus, if we would know the range of $[A, B*]^*$ to be weak* closed, then we could proceed as in the one-block case and write (DUAL') in the form (replace δ by $A^* \eta + B^* * \gamma$, and recall that $AH = a$ and $B * H = b$):

$$\begin{aligned} & \text{maximize } \nu = \langle a, \eta \rangle + \langle b, \gamma \rangle, \\ (\text{DUAL}) \quad & \mathbf{1}^* \tau = 1, \\ & -E^* \tau \leq A^* \eta + B^* * \gamma \leq E^* \tau. \end{aligned}$$

Indeed, as we shall prove in the following, if the matrices $\begin{pmatrix} \widehat{U}_0 \\ \widehat{U}_1 \end{pmatrix}$ and $\begin{pmatrix} \widehat{V}_0 & \widehat{V}_1 \end{pmatrix}$ have full rank on the unit circle, then the range of $[A, B*]^*$ will be weak* closed, but there

seems to be no reason to believe that the range of $[A, B^*]^*$ will be weak* closed in general. Still, even if the range of $[A, B^*]^*$ is not weak* closed, much of Lemma 5.3 and Theorem 5.4 remains true for the multiblock case.

We first prove a weakened analogue of Lemma 5.3.

LEMMA 6.1. *Define A , B , $[A, B^*]$, and $[A, B^*]^*$ as before. Then $[A, B^*]^*$ maps $\mathbf{R}^{\sigma_{UV}} \times c_{(mn-pq)}^0$ continuously into $c_{m \times n}^0$, and $[A, B^*]$ is the adjoint of the operator $[A, B^*]^*$ restricted to $\mathbf{R}^{\sigma_{UV}} \times c_{(mn-pq)}^0$. In particular, $\mathcal{N}([A, B^*])$ is weak* closed in $l_{m \times n}^1$, and (PRIM) is the dual of the problem (DUAL) posed in $\mathbf{R}^{\sigma_{UV}} \times c_{(mn-pq)}^0$.*

Proof. That $[A, B^*]^*$, which is equal to the operator $(\eta, \gamma) \mapsto A^*\eta + B^*\gamma$, maps $\mathbf{R}^{\sigma_{UV}} \times c_{(mn-pq)}^0$ continuously into $c_{m \times n}^0$ follows from the fact that A^* maps $\mathbf{R}^{\sigma_{UV}}$ into $c_{m \times n}^0$ and that B^* maps $c_{(mn-pq)}^0$ into $c_{m \times n}^0$ since $B \in l^1$. All other claims in Lemma 6.1 then follow immediately. \square

The preceding considerations can be collected into the following result.

THEOREM 6.2. *In the situation described in Theorem 4.4 (the multiblock case) problem (OPT) is equivalent to the infinite-dimensional linear programming problem (PRIM), and it has a minimizer $\Phi \in l^1$. The minimum is equal to the distance in l^1 of H to the weak*-closed null space of $[A, B^*]$. To compute this distance we may instead solve the dual problem (DUAL'), which has a maximizer $\delta \in l_{m \times n}^\infty$, or we may solve the problem (DUAL) in $\mathbf{R}^{\sigma_{UV}} \times c_{(mn-pq)}^0$, which need not have a maximizing solution, but for which the supremum of the objective function $\langle a, \eta \rangle + \langle b, \gamma \rangle$ over the set of all feasible solutions $(\eta, \gamma) \in \mathbf{R}^{\sigma_{UV}} \times c_{(mn-pq)}^0$ equals the optimal value of (PRIM). If, in addition, $\mathcal{R}([A, B^*])$ is closed in $\mathbf{R}^{\sigma_{UV}} \times l_{(mn-pq)}^1$, then (DUAL) has a maximizing solution in $\mathbf{R}^{\sigma_{UV}} \times l_{(mn-pq)}^\infty$.*

Here the final statement makes use of the closed range theorem, which says that the range of the dual of an operator is weak* closed if and only if the range of the operator itself is closed; see [12, Thm. 4.14, p. 96]. Theorem 6.2 extends the corresponding ones in [5] and [10], where the existence of a maximizer of the dual problem in l^∞ is never addressed.

For later reference, let us mention the following result, related to Lemme 6.1.

LEMMA 6.3. *Let Assumption 1.1 hold. Then the mapping that takes $Q \in l_{p \times q}^1$ into $K = U * Q * V$ and its inverse are both norm-continuous and weak* continuous.*

Proof. We know that these mappings are norm-continuous; see Theorem 4.4. The weak* continuity is a consequence of the fact that the mapping $Q \mapsto U * Q * V$ is the adjoint of a continuous operator mapping $c_{m \times n}^0$ into $c_{p \times q}^0$. \square

The additional condition on the range of $[A, B^*]$ in Theorem 6.2 is satisfied in the following case.

LEMMA 6.4. *If the matrices $(\hat{U}_0 \ \hat{U}_1)$ and $(\hat{V}_0 \ \hat{V}_1)$ have full rank on the unit circle, then the range of the operator $[A, B^*]$ is closed in $\mathbf{R}^{\sigma_{UV}} \times l_{(mn-pq)}^1$.*

Proof. Recall that, by the closed range theorem [12, Thm. 4.14], $\mathcal{R}([A, B^*])$ is closed in $\mathbf{R}^{\sigma_{UV}} \times l_{(mn-pq)}^1$ if and only if $\mathcal{R}([A, B^*]^*)$ is closed in $l_{m \times n}^\infty$. The range of the latter operator is the sum of the ranges of A^* and B^* . Since the range of A is finite-dimensional, to prove that $\mathcal{R}([A, B^*]^*)$ is closed, it suffices to prove that $\mathcal{R}(B^*)$ is closed. However, again by the closed range theorem, this is true if and only if $\mathcal{R}(B)$ is closed in $l_{(mn-pq)}^1$. Thus, we have reduced the proof to a proof of the fact that $\mathcal{R}(B)$ is closed in $l_{(mn-pq)}^1$.

To prove that $\mathcal{R}(B)$ is closed we shall decompose B into three operators, $B = PDC$, where C has a closed range, and D and P have the property that the images

of closed subspaces under these operators are closed. Clearly, this implies that $\mathcal{R}(B^*)$ is closed. This decomposition is based on the fact that it is possible to rewrite (17) into the form

$$(30) \quad \begin{pmatrix} \widehat{U}_0 \\ \widehat{U}_1 \end{pmatrix} K \begin{pmatrix} \widehat{V}_0 & \widehat{V}_1 \end{pmatrix} - \begin{pmatrix} I & 0 \\ \widehat{U}_1 \widehat{U} [\widehat{U}_0 \widehat{U}]^{-1} & 0 \end{pmatrix} \begin{pmatrix} \widehat{U}_0 \\ \widehat{U}_1 \end{pmatrix} K \begin{pmatrix} \widehat{V}_0 & \widehat{V}_1 \end{pmatrix} \begin{pmatrix} I & [\widehat{V} \widehat{V}_0]^{-1} \widehat{V} \widehat{V}_1 \\ 0 & 0 \end{pmatrix} = 0.$$

We define the operator C to be the operator that maps $K \in l_{m \times n}^\infty$ into

$$CK = \begin{pmatrix} U_0 \\ U_1 \end{pmatrix} * K * \begin{pmatrix} V_0 & V_1 \end{pmatrix} = \begin{pmatrix} U_0 * K * V_0 & U_0 * K * V_1 \\ U_1 * K * V_0 & U_1 * K * V_1 \end{pmatrix}.$$

Then, by Theorem 3.1, Lemma 5.1, and Lemma 5.3, this operator has closed range (of finite codimension).

To define the remaining operators D and P we have to partition $K \in l_{m \times n}^1$ into $K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$ conformally with the partition above. Recall that \widehat{U}_0 , \widehat{U}_1 , \widehat{V}_0 , and \widehat{V}_1 have been chosen in such a way that the functions $\widehat{U}_1 \widehat{U} [\widehat{U}_0 \widehat{U}]^{-1}$ and $[\widehat{V} \widehat{V}_0]^{-1} \widehat{V} \widehat{V}_1$ in (17) belong to \hat{l}^1 . This means that they are \mathcal{Z} -transforms of some functions in l^1 . Let us denote these l^1 -functions by X_0 and X_1 , respectively. Then B^* can be written in the form $B^* = PDC$, where C is the operator defined above, and D and P are the operators

$$DK = \begin{pmatrix} K_{11} & K_{12} - K_{11} * X_1 \\ K_{21} - X_0 * K_{11} & K_{22} - X_0 * K_{11} * X_1 \end{pmatrix}, \quad PK = \begin{pmatrix} 0 & K_{12} \\ K_{21} & K_{22} \end{pmatrix}.$$

The operator D has a continuous inverse, which we get from the preceding formula for D by changing all minus signs to plus signs. The operator P is a continuous projection operator. Both of these operators have the property that the image of a closed subspace is closed. Thus, $B^* = PDC$ has a closed range, as claimed. \square

7. Alignment conditions. Later we shall need the following well-known connections between the primal and the dual problems. For simplicity we formulate this only for the multiblock case with the dual problem expressed in the form (DUAL'), and we leave the one-block case and the case where the dual problem has been phrased in the form (DUAL) to the reader.

THEOREM 7.1. (i) If (μ, Φ^+, Φ^-) is a feasible solution of (PRIM) and (ν, τ, δ) is a feasible solution of (DUAL') (i.e., the constraints are satisfied, but not necessarily the minimality or maximality condition), then $\mu \geq \nu$.

(ii) If, in addition, $\mu = \nu$, then both solutions are optimal for their respective problem (i.e., the minimality and maximality conditions are satisfied).

(iii) In the optimal case the following alignment conditions are satisfied:

(a) if for some $i \in \{1, \dots, m\}$, $\sum_{jk} (\Phi_{ij}^+(k) + \Phi_{ij}^-(k)) < \mu$, then $\tau_i = 0$;

(b) if $\Phi_{ij}^+(k) > 0$ for some $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$, and $k \in \{0, 1, \dots\}$, then $\delta_{ij}(k) = \tau_i$;

(c) if $\Phi_{ij}^-(k) > 0$ for some $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$, and $k \in \{0, 1, \dots\}$, then $\delta_{ij}(k) = -\tau_i$;

(d) if for some $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$, and $k \in \{0, 1, \dots\}$, both $\Phi_{ij}^+(k) > 0$ and $\Phi_{ij}^-(k) > 0$, then $\tau_i = 0$.

Proof. Using the feasibility of the primal and dual solutions we get

$$\begin{aligned}
 \nu &= \langle H, \delta \rangle \\
 &= \langle \Phi^+ - \Phi^- - (\Phi^+ - \Phi^- - H), \delta \rangle \\
 &= \langle \Phi^+, \delta \rangle - \langle \Phi^-, \delta \rangle \\
 (31) \quad &\leq \langle \Phi^+, E^* \tau \rangle + \langle \Phi^-, E^* \tau \rangle \\
 &= \langle E(\Phi^+ + \Phi^-), \tau \rangle \\
 &\leq \mu \langle \mathbf{1}, \tau \rangle \\
 &= \mu.
 \end{aligned}$$

This proves part (i). Part (ii) is a direct consequence of part (i). Finally, by inspecting the inequalities in the computation above, and checking when they reduce to equalities, we get (iii). \square

If $\tau_i = 0$ for some $i \in \{1, \dots, m\}$, then the corresponding double inequalities in (DUAL) and (DUAL') collapse into equalities. This means that there is an extensive degeneracy in the dual problem, and that we can expect an infinite-dimensional nonuniqueness in the primal problem. It follows from (a) and (d) in Theorem 7.1 that this happens whenever the optimal primal solution contains a row i satisfying $\sum_{jk} |\Phi_{ij}(k)| < \mu$. Of course, if $m = 1$, i.e. U contains only one row, then this problem never shows up, since in that case $\tau_1 = 1$. For more details, see the discussion in §§8 and 9.

8. Solution methods. In [5] Dahleh and Pearson introduced a method for solving the linear programming version of (OPT), and the same method was used in [10]. In this method we restrict the number of nonzero variables $\Phi(k)$ in the primal problem, i.e., we solve a sequence of truncated problems

$$\begin{aligned}
 &\text{minimize } \mu, \\
 &1\mu - E(\Phi^+ + \Phi^-) \geq 0, \\
 &A(\Phi^+ - \Phi^-) = a, \\
 (\text{PRIM})_N \quad &B * (\Phi^+ - \Phi^-) = b, \\
 &(\Phi^+)(k) = (\Phi^-)(k) = 0, \quad k > N, \\
 &\Phi^+, \Phi^- \geq 0,
 \end{aligned}$$

where $N \rightarrow \infty$. The dual of this problem is

$$\begin{aligned}
 &\text{maximize } \nu = \langle a, \eta \rangle + \langle b, \gamma \rangle, \\
 (\text{DUAL})_N \quad &1^* \tau = 1, \\
 &-E^* \tau \leq (A^* \eta + B^* * \gamma)(k) \leq E^* \tau, \quad k \in \{0, 1, \dots, N\}.
 \end{aligned}$$

Clearly, since $(\text{PRIM})_N$ contains additional restrictions compared to (PRIM), and since the number of these additional restrictions decrease for increasing N , the optima μ_N of the truncated problems form a nonincreasing sequence, bounded from below by the optimum μ of (PRIM). Moreover, since the set of sequences with finite support is dense in l^1 , it is clear that $\mu_N \rightarrow \mu$ as $N \rightarrow \infty$. We shall refer to this method as the FMV method (finitely many variables in the primal problem).

As it was pointed out in [5] and [10], the FMV method has two drawbacks. One of them is that it cannot always be applied, due to the fact that the systems $(\text{PRIM})_N$

need not have any feasible solutions, i.e., there are cases where (PRIM) has no feasible solution Φ with finite support. A precise description of this class of problems where the FMV method fails was given in [10]; in particular, it requires \widehat{H} , \widehat{U} , and \widehat{V} to be rational functions. Another serious drawback is that there is no way of telling how close the optimum μ_N of the truncated problem (PRIM) $_N$ is to the optimum μ of the full problem, i.e., there is no way of knowing how good the solution of the truncated problem is.

The drawbacks mentioned previously prompted the author to propose another method in [14] for the solution of (PRIM); the same method was proposed independently by Dahleh in [3]. In this approach we drop all but finitely many equations, i.e., we solve the problem

$$\begin{aligned}
 & \text{minimize } \mu, \\
 (\text{PRIM})_M \quad & \mathbf{1}\mu - E(\Phi^+ + \Phi^-) \geq 0, \\
 & A(\Phi^+ - \Phi^-) = a, \\
 & (B * (\Phi^+ - \Phi^-))(\ell) = b(\ell), \quad \ell \in \{0, 1, \dots, M\}, \\
 & \Phi^+, \Phi^- \geq 0,
 \end{aligned}$$

where the constraints $(B * \Phi)(\ell) = b(\ell)$ have been ignored for $\ell > M$ for some number M , and let $M \rightarrow \infty$. The dual of a primal problem truncated in this way is a truncated version of (DUAL),

$$\begin{aligned}
 & \text{maximize } \nu = \langle a, \eta \rangle + \langle b, \gamma \rangle, \\
 (\text{DUAL})_M \quad & \mathbf{1}^* \tau = 1, \\
 & -E^* \tau \leq (A^* \eta + B^* * \gamma) \leq E^* \tau, \\
 & \gamma(\ell) = 0, \quad \ell > M,
 \end{aligned}$$

where we impose the extra condition $\gamma(\ell) = 0$ for $\ell > M$. Since (DUAL) $_M$ contains additional restrictions compared to (DUAL), and since the number of these additional restrictions decrease for increasing M , it is clear that the optima of the truncated problems form a nondecreasing sequence, bounded from above by the supremum of (DUAL) posed in $\mathbf{R}^{\sigma_{UV}} \times c^0_{(mn-pq)}$. Moreover, since the set of sequences with finite support is dense in c^0 , the optima of the truncated dual problem tend to the supremum of (DUAL) in $\mathbf{R}^{\sigma_{UV}} \times c^0_{(mn-pq)}$ as $M \rightarrow \infty$. We shall refer to this method as the FME method (finitely many equations in the primal problem).

Since the optimum of (PRIM) is equal to the supremum of (DUAL) in $\mathbf{R}^{\sigma_{UV}} \times c^0_{(mn-pq)}$, the limits of the optima obtained from the FMV method and the FME method are the same. Thus, by applying both these methods at the same time we get both upper bounds and lower bounds on the true optimum, and we may compute the optimum of the full problem to within any given tolerance.

Although it is clear in principle what we should do, one major obstacle remains: How do we solve the truncated problems?

Let us first discuss the solution of the truncated problems in the FMV method. For this method to be applicable we have to assume that the tensor sequence B and the vector sequence b have finite support. Then, if we require $\Phi(k) = 0$ for $k > N$, then both $(B * \Phi)(\ell)$ and $b(\ell)$ vanish for sufficiently large ℓ . This means that, although the problem formally is infinite dimensional, all but finitely many equations are satisfied trivially. Thus, instead of having an infinite-dimensional linear

programming system we have a finite-dimensional one, to which we may apply any standard linear programming solution method. (It may be the case that this finite-dimensional system has no feasible solution for any N , and in this case the FMV method fails; however, in a fair number of cases feasible solutions do exist; cf. [10].)

The situation is more complicated in the FME case, because there it is not obvious that each truncated system can be reduced to a finite-dimensional one. (As we shall see in a moment, this is indeed the case.) However, there is one simple case where everything is very straightforward, namely, the case where \mathcal{Z}_{UV} is empty. In this case the operator A is not present in (PRIM) and (DUAL). If we force $\gamma(\ell) = 0$ for $\ell > M$, then, by (29), $B^* \cdot \gamma(k) = 0$ for $k > M$. Thus, there are only a finite number of inequalities in (DUAL) that are not satisfied trivially (recall that $(E^* \tau)_{ij} = \tau_i \geq 0$). This means that we are again left with a finite-dimensional linear programming problem that can be solved with standard methods. This special case seems to be of substantial interest as many real world problems can be formulate in this way; see [13].

The argument above does not apply to a typical one-block case or multiblock case where A is present (if A is absent in the one-block case then the optimal solution is 0; the solution of the multiblock case may be nontrivial even if A is zero). To understand the multiblock case it is crucial that we understand how the one-block case may be handled; therefore let us first discuss the one-block case, where the convolution operators B^* and B^{**} are absent. That case was solved in [4] in the special case where only the zero-order derivatives are needed in (14). The solution of the general case is based on the following lemma (cf. [4, Thm. 5]).

LEMMA 8.1. *Let C be an operator from \mathbf{R}^r to one-dimensional real c^0 . Then there is some $N \geq 0$ such that $\sup_{k > N} |(C\eta)(k)| \leq \max_{k \leq N} |(C\eta)(k)|$ for every $\eta \in \mathbf{R}^r$, and the inequality is strict whenever $C\eta \neq 0$.*

Proof. For each k , we may write $(C\eta)(k)$ in the form $(C\eta)(k) = \sum_{j=1}^r c_{kj}\eta_j$. Thus we may regard C as an $\infty \times r$ -dimensional matrix. The assumption that C maps \mathbf{R}^r into c^0 means that $c_{kj} \rightarrow 0$ as $k \rightarrow \infty$.

Without loss of generality, we may suppose that the rank of C is r (otherwise the column rank of C will be less than r , and we may drop some of the columns of C without affecting the range of C , and decrease r). We may further assume that the mapping which takes η into the first r components of $C\eta$ has rank r (otherwise permute some of the rows of C). Let x represent the first r components of $C\eta$, and let D be the matrix of rank r that maps η into x (D consist of the first r rows of C). Then $\eta = D^{-1}x$.

For each k we have $|(C\eta)(k)| \leq \sum_{j=1}^r |c_{kj}\eta_j| \leq \|C_k\|_1 \|\eta\|_\infty$, where $\|C_k\|_1 = \sum_j |c_{kj}|$ and $\|\eta\|_\infty = \max_j |\eta_j|$. Replacing η by $D^{-1}x$, we get $|(C\eta)(k)| \leq \|C_k\|_1 \|D^{-1}\| \|x\|_\infty$, where $\|D^{-1}\|$ is the operator norm of D^{-1} as an operator from \mathbf{R}^r with the l^∞ -norm to itself. Thus, if we choose N so large that $\|C_k\|_1 \|D^{-1}\| < 1$ for all $k > N$, then

$$\sup_{k > N} |(C\eta)(k)| \leq \|C_k\|_1 \|D^{-1}\| \|x\|_\infty \leq \|x\|_\infty = \max_{k \leq r} |(C\eta)(k)| \leq \max_{k \leq N} |(C\eta)(k)|,$$

and the inequality is strict whenever $x \neq 0$, i.e., $C\eta \neq 0$. \square

By using Lemma 8.1 it is easy to prove the following theorem for the one-block case.

THEOREM 8.2. *In the one-block case the dual problem (DUAL₁) is equivalent to a problem with finitely many constraints, i.e., there is some $N > 0$ such that it is*

equivalent to the problem

$$\begin{aligned}
 & \text{maximize } \nu = \langle a, \eta \rangle, \\
 (\text{DUAL}_1)_N \quad & \mathbf{1}^* \tau = 1, \\
 & -E^* \tau \leq (A^* \eta)(k) \leq E^* \tau \quad k \in \{0, 1, \dots, N\}.
 \end{aligned}$$

This number N depends only on the zero structures of \widehat{U} and \widehat{V} ; in particular, it is independent of H .

Proof. For each i , let D_i be the operator that maps $\eta \in \mathbf{R}^{\sigma_{UV}}$ into $(A^* \eta)_{ij}(k)$. As it was pointed out when the operator A was first introduced, $A(k) \rightarrow 0$ as $k \rightarrow \infty$; hence D_i maps $\mathbf{R}^{\sigma_{UV}}$ into c^0 . Apply Lemma 8.1 to conclude that there is some number N_i such that $\sup_{1 \leq j \leq n, k > N} |(A^* \eta)_{ij}(k)| \leq \max_{1 \leq j \leq n, k \leq N} |(A^* \eta)_{ij}(k)|$ for every $\eta \in \mathbf{R}^{\sigma_{UV}}$. Choose $N = \max\{N_i\}$ to get the conclusion of Theorem 8.2. \square

COROLLARY 8.3. *In the one-block case the primal problem (PRIM_1) has a solution Φ that satisfies $\Phi(k) = 0$ for $k > N$, where N is the number given by Theorem 8.2. Moreover, it is possible to choose Φ in such a way that at most $\sigma_{UV} + m - 1$ of the components $\Phi_{ij}(k)$ are nonzero.*

The first claim follows from the fact that the dual problem of $(\text{DUAL}_1)_N$ is equal to (PRIM_1) with the additional restriction $\Phi^+(k) = \Phi^-(k) = 0$ for $k > N$, and the second claim is a standard result saying that in a system with $\sigma_{UV} + m$ inequalities we can find a solution with at most $\sigma_{UV} + m$ nonzero components, and one of these nonzero components is the variable μ . This corollary is closely related to [10, Thm. 5], although the proof given here is quite different from the one in [10].

The same argument can be applied to problem $(\text{DUAL})_M$, since (29) implies that $B^* * \gamma(k) \rightarrow 0$ as $k \rightarrow \infty$ whenever γ has finite support. Thus, the following theorem is true.

THEOREM 8.4. *In the multiblock case the problem $(\text{DUAL})_M$ is equivalent to a problem with finitely many constraints, i.e., there is some $N > 0$ such that it is equivalent to the problem*

$$\begin{aligned}
 & \text{maximize } \nu = \langle a, \eta \rangle + \langle b, \gamma \rangle, \\
 (\text{DUAL})_{MN} \quad & \mathbf{1}^* \tau = 1, \\
 & -E^* \tau \leq (A^* \eta + B^* * \gamma)(k) \leq E^* \tau, \quad k \in \{0, 1, \dots, N\}, \\
 & \gamma(\ell) = 0, \quad \ell > M.
 \end{aligned}$$

This number N is independent of H .

This means that the FME method also amounts to the solution of a sequence of finite-dimensional linear programming problems.

The problem $(\text{DUAL})_{MN}$ can be considered as the dual of the following problem:

$$\begin{aligned}
 & \text{minimize } \mu, \\
 (\text{PRIM})_{MN} \quad & \mathbf{1}\mu - E(\Phi^+ + \Phi^-) \geq 0, \\
 & A(\Phi^+ - \Phi^-) = a, \\
 & (B * (\Phi^+ - \Phi^-))(\ell) = b(\ell), \quad \ell \in \{0, 1, \dots, M\}, \\
 & (\Phi^+)(k) = (\Phi^-)(k) = 0, \quad k > N, \\
 & \Phi^+, \Phi^- \geq 0.
 \end{aligned}$$

Observe that this is exactly the same problem that we solve in the FMV method, except for the fact that in the FMV method M is large compared to N , whereas

in the FME method N is large compared to M . Since all standard programs for solving finite-dimensional linear programming systems produce the solution to both the primal problem and the dual problem at the same time, it does not really matter which one of the two problems $(\text{PRIM})_{MN}$ or $(\text{DUAL})_{MN}$ that we solve. This means that the only essential difference in the two methods is the difference in the sizes of M and N .

9. Determining the size of a truncated problem. According to the discussion in the preceding section, the FME method can always be applied, even when the functions \hat{H} , \hat{U} , and \hat{V} are not rational, and the sequence of optima that we get for the truncated problems converges to the optimum of the original problem. However, this is not the whole truth. Theorem 8.4 does not give any explicit formula for the calculation of the number N . A very crude estimate can be obtained from the proofs of Lemmas 8.1 and 8.2: At each step (i.e., for each M) we invert m different square matrices of size $(\sigma_{UV} + (mn - pq)(M + 1))^2$ (one for each i), and compute the l^∞ operator norms of these matrices. (We believe that it is possible to get a much better estimate based on the replacement of the l^∞ matrix norm by the l^2 matrix norm, but even the computation of this improved estimate requires a fair amount of work.) In the one-block this phase appears only once, and it can be afforded, but in the multiblock case this produces becomes rather expensive in terms of computation time. In addition, there is no guarantee that the problems do not become numerically ill behaved with increasing M . For this reason it would be very desirable to have some alternative method of estimating the number N corresponding to the number M in the FME method. (The upper bound on M that corresponds to a given N in the FMV method is of the type $M = N + \text{constant}$, so the computation of M in the FMV method presents no problems.)

In the case $m = 1$ there is a simple and effective solution to this problem. Recall that when $m = 1$, we have $\tau_1 = 1$. Moreover, recall that $(B^* * \gamma)(k) = 0$ for $k > M$. Thus, the problem is to determine, for each M and η , a bound $N > M$ such that

$$\left| \sum_{r=1}^{\sigma_{UV}} \eta_r A_{r1j}(k) \right| \leq 1, \quad k > N.$$

However, this is easy. If we use a standard Simplex type solution method, then at each stage of the computation we have explicit access to the vector η that currently is considered, and it suffices to choose $N > M$ so large that

$$\sup_r |\eta_r| \sum_{r=1}^{\sigma_{UV}} |A_{r1j}(k)| \leq 1, \quad k > N.$$

Thus, it is possible to determine during each step of the solution process exactly how large values of k need to be considered at that moment, and we can avoid the costly computation of the a priori upper bounds on N . If necessary, it is not difficult to increase the size of N during the solution process, since the matrix A has a form that makes it easy to compute additional coefficients from the original zero structures of \hat{U} and \hat{V} .

The case where $m > 1$ is more problematic. The same method that we just mentioned works in the case where each $\tau_i > 0$; the only difference is that we have to take $N > M$ so large that, for all $i \in \{1, \dots, m\}$ and all $j \in \{1, \dots, n\}$,

$$\sup_r |\eta_r| \sum_{r=1}^{\sigma_{UV}} |A_{rij}(k)| \leq \tau_i, \quad k > N.$$

If the optimal solution of (DUAL) is such that $\tau_i > 0$ for all i , then we expect the intermediate problems to have the same property for large enough M , and the computations can be carried out in an efficient way. (For further comments on this, see §§10 and 11.) If, however, at some stage of the computation one or more of the variables τ_i become zero, then we must use some alternative method, such as the one outlined in the beginning of this section.

The problem of having one or more of the variables $\tau_i = 0$ is more severe than it might appear at first glance, because it is quite common that the optimal solution has some of the variables τ_i equal to zero. See the discussion following Theorem 7.1. One possible solution to this problem is based on the following theorem that was proved in [15, Thm. 5.2] (that theorem is stated in a slightly different context, but the same proof applies without change).

THEOREM 9.1. *The matrix sequence that we get from an optimal solution Φ of the primal problem by dropping those rows i for which*

$$\sum_{j,k} |\Phi_{ij}(k)| < \max_i \sum_{j,k} |\Phi_{ij}(k)|$$

is optimal for the corresponding reduced problem, i.e., the problem that we get by dropping the corresponding rows of H and U . Moreover, the optimal values of the full problem and the reduced problem are the same.

Thus, if it appears that there are rows i in the optimal solution for which

$$\sum_{j,k} |\Phi_{ij}(k)| < \max_i \sum_{j,k} |\Phi_{ij}(k)|,$$

then one possibility is to simply drop these rows, and to solve a smaller problem.

There is a very interesting open question related to this approach: Is it always true that the smaller problem has the same rank as the original problem, and that it satisfies Assumption 1.1? (We conjecture that the answer is yes, at least generically.) If this is the case, then it is possible to construct a continuous, one-to-one mapping between all the feasible solutions of the smaller problem and the feasible solutions of the full problem.

LEMMA 9.2. *Let $I \subset \{1, \dots, m\}$ be an index set, and let $(\text{OPT})_I$ be the problem that we get from (OPT) by dropping all those rows i of H and U for which $i \notin I$. Suppose that the new matrix U has the same rank p as the original matrix has (in particular, I contains at least p indices; hence $m > p$, and we are in the multiblock case), and that the new matrix U satisfies Assumption 1.1. Then there is a bicontinuous one-to-one correspondence between the feasible solutions of $(\text{OPT})_I$ and the feasible solutions of (OPT).*

Proof. Let P_I be the $m \times m$ -dimensional projection matrix

$$(32) \quad P_I = \text{diag}(d(i)), \text{ where } d(i) = \begin{cases} 1, & \text{if } i \in I, \\ 0, & \text{if } i \notin I. \end{cases}$$

Then the reduced problem $(\text{OPT})_I$ can be written in the form

$$(\text{OPT})_I \quad \text{minimize } \|\Phi_I\|_{l^1} \text{ over all } \Phi_I \text{ of the form } \Phi_I = P_I H - P_I U * Q * V \text{ where } Q \in l^1 \text{ and } \|\Phi_I\|_{l^1} \text{ is given by (2).}$$

Clearly, each feasible solution Φ of (OPT) induces a feasible solution $\Phi_I = P_I \Phi$ of $(\text{OPT})_I$, and the mapping from Φ to Φ_I is continuous.

Conversely, suppose that $P_I U$ has rank p , and that Assumption 1.1 holds with \hat{U} replaced by $P_I \hat{U}$. Let Φ_I be a feasible solution of $(\text{OPT})_I$. By Lemma 6.3, there is a continuous map from Φ_I onto the free parameter Q in $(\text{OPT})_I$. Let Φ be the function that we get from Φ_I by first mapping $\Phi_I = P_I H - P_I U * Q * V$ onto Q , and then mapping Q onto $\Phi = H - U * Q * V$. Then Φ depends continuously on Φ_I , Φ is a feasible solution of (OPT) , and $\Phi_I = P_I \Phi$, as required. \square

Another possibility in the degenerate case is to perturb the original problem so that, instead of using the pure operator-norm $\|\Phi\|_{l^1}$ of Φ given by (2), we use the norm

$$(33) \quad \|\Phi\|_\epsilon = \max_{i \in \{1, \dots, m\}} \sum_{j=1}^n \sum_{k=0}^{\infty} |\Phi_{ij}(k)| + \epsilon \sum_{i=1}^m \sum_{j=1}^n \sum_{k=0}^{\infty} |\Phi_{ij}(k)|,$$

where ϵ is some small positive constant. When this change is incorporated in (DUAL) , then the crucial bound (9) is relaxed to

$$(34) \quad \sup_r |\eta_r| \sum_{r=1}^{\sigma_{UV}} |A_{rij}(k)| \leq \tau_i + \epsilon, \quad k > N.$$

Here the right-hand side cannot be zero, and the method described above works.

10. Convergence of primal and dual variables: FMV scheme. So far we have discussed only the convergence of the optima of the truncated problems to the true optimum. Two other interesting questions are whether the full solutions of the truncated primal problems converge to a full solution of the original primal problem, and whether the full solutions of the truncated dual problems tend to a full solution of the original dual problem. Apart from their theoretical interest, these questions have some practical consequences. One major issue on the dual side is that if the determination of the cutoff number N in the FME scheme is carried out as proposed in §9 then, for computational reasons, it is vital that the weights τ_i corresponding to the optimal solutions of the truncated dual problems behave in a consistent way. More specifically, they should stay bounded away from zero except in the case where the corresponding row is redundant in the optimal solution. In the latter case they should tend to zero. Another possible issue on the primal side, the importance of which is difficult to judge at this time due to the immature state of the theory, is the following. A typical approximate solution produced by the linear programming method is of very high order, and the order is growing without bound as we approach the true optimum. In many case the order of the approximate solution is so high that it does not make sense to implement this particular solution in a design. However, as some evidence presented in [14] and [15] indicates, an optimal solution may often be significantly “simpler,” e.g., of much lower order, than the approximate solutions. Suppose, indeed, that we can prove that the approximate solutions are “close” to a true optimal solution. Then it should be possible to use some order reduction scheme to recover a low order near optimal solution.

In this section we look first at the problem of the convergence of the complete solution of the truncated linear programming problems in the FMV scheme. The same question for the FME scheme is discussed in §11.

We begin with the question of the convergence of the primal variables in the FMV scheme.

THEOREM 10.1. (i) *For the FMV scheme, each subsequence of a sequence Φ_N of optimal solutions to the problem $(\text{PRIM})_N$ contains a subsequence that converges*

weak* in $l^1_{m \times n}$ to an optimal solution Φ of (PRIM). If, moreover, the optimal solution of (PRIM) is unique, then the whole sequence Φ_N tends to the optimal solution Φ of (PRIM) weak* in $l^1_{m \times n}$.

(ii) Let Φ be a optimal solution of (PRIM) that is a weak* limit of some subsequence Φ_{N_ℓ} of optimal solutions Φ_N of (PRIM) $_N$, as described in (i). Let I denote the set of indices $i \in \{1, \dots, m\}$ for which $\sum_{jk} |\Phi_{ij}(k)| = \mu = \sup_i \sum_{jk} |\Phi_{ij}(k)|$, and define the projection matrix P_I as in (32). Then $P_I \Phi_{N_\ell}$ tends to $P_I \Phi$ in the norm of $l^1_{m \times n}$. In other words, if we delete those rows of Φ_{N_ℓ} for which $\sum_{jk} |\Phi_{ij}(k)| < \mu = \sup_i \sum_{jk} |\Phi_{ij}(k)|$, then the remaining part of Φ_{N_ℓ} tends to the corresponding part of Φ in the norm of $l^1_{m \times n}$.

(iii) Let P_I be the projection operator defined in (ii). Suppose that $P_I U$ has the same rank as U , and that Assumption 1.1 holds with \hat{U} replaced by $P_I \hat{U}$. Then the subsequence Φ_{N_ℓ} in (ii) tends to Φ in the norm of $l^1_{m \times n}$.

Compare part (ii) of this theorem to Theorem 9.1. In both places the set of exceptional indices is the same.

Proof. (i) Since the norms $\mu_N = \|\Phi_N\|_{l^1}$ form a bounded (monotonically non-increasing) sequence, the sequence Φ_N is bounded in l^1 . Bounded subsets of l^1 are weak*-sequentially compact; hence each subsequence contains a subsequence that tends weak* to some limit Φ . The limit Φ must satisfy $\Phi - H \in \mathcal{N}([A, B*])$, since this null space is weak* closed, and since $\Phi_N - H \in \mathcal{N}([A, B*])$ for all N . Thus, Φ is an optimal solution of (PRIM). The final conclusion in (i) follows from the fact that if every subsequence has a subsequence that converges to the same limit, then the whole sequence must converge to this limit.

(ii) To simplify the notations slightly, let us replace the subsequence Φ_{N_ℓ} by the original sequence Φ_N . By Fatou's lemma, we have for all $i \in \{1, \dots, m\}$,

$$\sum_{jk} |\Phi_{ij}(k)| \leq \liminf_{N \rightarrow \infty} \sum_{jk} |(\Phi_N)_{ij}(k)|.$$

On the other hand, since $\lim_{N \rightarrow \infty} \|\Phi_N\|_{l^1} = \lim_{N \rightarrow \infty} \mu_N = \mu = \|\Phi\|_{l^1}$, where μ is the optimal value of (PRIM), the reverse inequality holds for all $i \in I$ with \liminf replaced by \limsup . Thus

$$\sum_{jk} |\Phi_{ij}(k)| = \lim_{N \rightarrow \infty} \sum_{jk} |(\Phi_N)_{ij}(k)|, \quad i \in I.$$

However, weak* convergence together with convergence of the norm implies norm convergence in one-dimensional l^1 (the norm in one-dimensional c^0 has the Kadec-Klee property), and we conclude that $P_I \Phi_N$ tends to $P_I \Phi$ in the norm of l^1 , as claimed.

(iii) This follows from (ii) and Lemma 9.2. \square

A partial answer to the uniqueness problem for (PRIM) is given in the following theorem.

THEOREM 10.2. *Let (ν, τ, δ) be an optimal solution of (DUAL'), and define the sets I and I' of indices $i \in \{1, \dots, m\}$ and the sets J , J^+ , and J^- of indices*

$i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$, and $k \in \{0, 1, \dots\}$ as follows:

$$\begin{aligned} I &= \{i \mid \tau_i > 0\}, \\ I' &= \{i \mid \tau_i = 0\}, \\ J &= \{(i, j, k) \mid -\tau_i < \delta_{ij}(k) < \tau_i\}, \\ J^+ &= \{(i, j, k) \mid \delta_{ij}(k) = \tau_i > 0\}, \\ J^- &= \{(i, j, k) \mid \delta_{ij}(k) = -\tau_i < 0\}. \end{aligned}$$

Then the optimal solution (μ, Φ) of (PRIM) is unique (under the additional normalizing assumption that for all i, j , and k , either $\Phi_{ij}^+(k) = 0$ or $\Phi_{ij}^-(k) = 0$) if and only if the following conditions determine $\mu \in \mathbf{R}$ and $\Phi \in l_{m \times n}^1$ uniquely:

$$\begin{aligned} A\Phi &= a, \\ B * \Phi &= b, \\ \sum_{jk} |\Phi_{ij}(k)| &= \mu, & i \in I, \\ \sum_{jk} |\Phi_{ij}(k)| &\leq \mu, & i \in I', \\ \Phi_{ij}(k) &= 0, & (i, j, k) \in J, \\ \Phi_{ij}(k) &\geq 0, & (i, j, k) \in J^+, \\ \Phi_{ij}(k) &\leq 0, & (i, j, k) \in J^-. \end{aligned} \tag{35}$$

Without the normalizing assumption that for all i, j , and k , either $\Phi_{ij}^+(k) = 0$ or $\Phi_{ij}^-(k) = 0$, the solution will not, in general, be unique. This relationship will be satisfied for all indices $i \in I$, but not necessarily for all $i \in I'$.

Proof. By Theorem 7.1, an optimal solution Φ must satisfy (35). Conversely, if Φ satisfies (35), then it is feasible, and checking the computation (31) we find that Φ must be optimal. Thus, every optimal solution satisfies (35), and every solution of (35) is optimal. \square

One interesting fact about Theorem 10.2 is that the conditions (35) do not refer to the dual solution directly; it only refers to the index sets I, I', J, J^+ , and J^- . This fact made it possible to construct the exact solutions of some two-block problems in [14]–[15]. In these cases it was possible to conjecture from the computer printouts what the correct index sets ought to be, and subsequently it was possible to use the equations in (35) (the inequalities in (35) turned out to be redundant) to construct exact optimal solutions. A proof of the optimality was also given, based on the construction of dual solutions corresponding to these index sets.

At this time it is not clear to what extent we may expect the dual solutions of $(\text{DUAL})_N$ to converge to a dual solution of (DUAL). If they stay bounded, then some subsequence must converge weak*, but we do not know what type of conditions that we would have to impose to guarantee the boundedness of a sequence of dual solutions.

11. Convergence of primal and dual variables: FME scheme. Much of Theorem 10.1 remains true for the FME scheme.

THEOREM 11.1. (i) *For the FME scheme, each subsequence of a sequence Φ_M of optimal solutions to the problem $(\text{PRIM})_M$ contains a subsequence that converges*

weak* in $l_{m \times n}^1$ to an optimal solution Φ of (PRIM). If, moreover, the optimal solution of (PRIM) is unique, then the whole sequence Φ_M tends to the optimal solution Φ of (PRIM) weak* in $l_{m \times n}^1$.

(ii) Let Φ be an optimal solution of (PRIM) that is a weak* limit of some subsequence Φ_{M_ℓ} of optimal solutions Φ_M of (PRIM) $_M$, as described in (i). Let I denote the set of indices $i \in \{1, \dots, m\}$ for which $\sum_{jk} |\Phi_{ij}(k)| = \mu = \sup_i \sum_{jk} |\Phi_{ij}(k)|$, and define the projection matrix P_I as in (32). Then $P_I \Phi_{M_\ell}$ tends to $P_I \Phi$ in the norm of $l_{m \times n}^1$. In other words, if we delete those rows of Φ_{M_ℓ} for which $\sum_{jk} |\Phi_{ij}(k)| < \mu = \sup_i \sum_{jk} |\Phi_{ij}(k)|$, then the remaining part of Φ_{M_ℓ} tends to the corresponding part of Φ in the norm of $l_{m \times n}^1$.

The proof of this theorem is essentially the same as the proof of Theorem 10.1.

There is one part missing in Theorem 11.1 compared to Theorem 10.1, namely, part (iii). Part (iii) does not carry over because the solutions of (PRIM) $_M$ are not feasible for the full problem (PRIM); hence, Lemma 9.2 cannot be applied.

The question of the convergence of the dual solutions is much easier for the FME scheme than for the FMV scheme.

THEOREM 11.2. (i) For the FME scheme, every subsequence of a sequence $(\nu_M, \tau_M, \eta_M, \gamma_M)$ of optimal solutions of the problem (DUAL) $_M$ contains a subsequence that tends to an optimal solution (ν, τ, δ) of (DUAL') in the sense that $\nu_{M_\ell} \rightarrow \nu$, $\tau_{M_\ell} \rightarrow \tau$ in \mathbf{R}^m , and $A^* \eta_{M_\ell} + B^* \gamma_{M_\ell} \rightarrow \delta$ weak* in $l_{m \times n}^\infty$. If, moreover, the solution of (DUAL') is unique, then the whole sequence $(\nu_M, \tau_M, \eta_M, \gamma_M)$ tends to (ν, τ, δ) in the sense described above.

(ii) If the matrices $\begin{pmatrix} \hat{V}_0 \\ \hat{V}_1 \end{pmatrix}$ and $\begin{pmatrix} \hat{V}_0 & \hat{V}_1 \end{pmatrix}$ have full rank on the unit circle, then there exists a bounded sequence $(\nu_M, \tau_M, \eta_M, \gamma_M)$ of optimal solutions of (DUAL) $_M$, and each subsequence of such a bounded sequence has a subsequence that tends to an optimal solution $(\nu, \tau, \eta, \gamma)$ of (DUAL) in the sense that $\nu_{M_\ell} \rightarrow \nu$, $\tau_{M_\ell} \rightarrow \tau$ in \mathbf{R}^m , $\eta_{M_\ell} \rightarrow \eta$ in $\mathbf{R}^{\sigma_{UV}}$, and $\gamma_{M_\ell} \rightarrow \gamma$ weak* in $l_{(mn-pq)}^\infty$.

(iii) If the matrices $\begin{pmatrix} \hat{V}_0 \\ \hat{V}_1 \end{pmatrix}$ and $\begin{pmatrix} \hat{V}_0 & \hat{V}_1 \end{pmatrix}$ have full rank everywhere in closed the unit disk \overline{D} , then the operator $[A, B]^*$ is one-to-one, and every sequence of optimal solutions $(\nu_M, \tau_M, \eta_M, \gamma_M)$ of (DUAL) $_M$ is bounded. Thus, in this case (ii) applies to all possible sequences of optimal solutions of (DUAL) $_M$.

(iv) The solution of (DUAL) is unique if and only if the operator $[A, B]^*$ is one-to-one and the solution of (DUAL') is unique. In this case every sequence of solutions $(\nu_M, \tau_M, \eta_M, \gamma_M)$ in (ii) is bounded and tends to the optimal solution $(\nu, \tau, \eta, \gamma)$ of (DUAL) in the sense described in (ii).

Proof. (i) By the restrictions $\mathbf{1}^* \tau_M = 1$ and $-E^* \tau_M \leq A^* \eta_M + B^* \gamma_M \leq E^* \tau_M$ in (DUAL) $_M$, the sequence $A^* \eta_M + B^* \gamma_M$ stays in the unit ball of $l_{m \times n}^\infty$; hence every subsequence of $(\nu_M, \tau_M, \eta_M, \gamma_M)$ contains a subsequence that converges in the sense described in (i) to a limit (ν, τ, δ) . We know from before that $\nu = \lim_{M \rightarrow \infty} \nu_M$ is the optimal value of (DUAL'). Clearly, the limit satisfies $\nu = \langle H, \delta \rangle$, $\mathbf{1}^* \tau = 1$ and $-E^* \tau \leq \delta \leq E^* \tau$. Moreover, $\delta \in \mathcal{N}([A, B]^*)^\perp$ since $A^* \eta_M + B^* \gamma_M \in \mathcal{N}([A, B]^*)^\perp$ for all M , and $\mathcal{N}([A, B]^*)^\perp$ is weak* closed. Thus, the limit is an optimal solution of (DUAL').

(ii) By Lemma 6.4 and the closed range theorem, the range of $[A, B]^*$ is closed. By the open mapping theorem, the mapping $[A, B]^*$ is open from $\mathbf{R}^{\sigma_{UV}} \times l_{(mn-pq)}^\infty$ onto its range. Thus, there is a bounded subset of $\mathbf{R}^{\sigma_{UV}} \times l_{(mn-pq)}^\infty$ that gets mapped onto the intersection of the unit ball in $l_{m \times n}^\infty$ with $\mathcal{R}([A, B]^*)$. Since $A^* \eta + B^* \gamma$ stays in this intersection whenever (η, γ) is part of a solution of (DUAL) $_M$, it is possible to

choose (η_M, γ_M) to belong to this bounded set. Thus, it is possible to find a bounded sequence of optimal solutions $(\nu_M, \tau_M, \eta_M, \gamma_M)$ of $(\text{DUAL})_M$. The rest of the proof is similar to the proof of (i).

(iii) The claim that $[A, B^*]^*$ is one-to-one is equivalent to the claim that $[A, B^*]$ is surjective (since the range of $[A, B^*]$ is closed). This claim is equivalent to the claim that B^* is surjective, and that A maps $\mathcal{N}(B^*)$ onto $\mathbf{R}^{\sigma_{UV}}$.

Since we assume that the matrices $\begin{pmatrix} \widehat{U}_0 \\ \widehat{U}_1 \end{pmatrix}$ and $\begin{pmatrix} \widehat{V}_0 & \widehat{V}_1 \end{pmatrix}$ have full rank everywhere in closed the unit disk \overline{D} , the operator C defined in the proof of Lemma 6.4 is surjective. The operators D and P defined in the same proof are also surjective. Thus, $B^* = PDC$ is surjective.

The null space of B^* consists of all functions that satisfy (16). By Lemma 4.2, it can be parametrized as the set of all functions K that satisfy $\widehat{K} = \widehat{L}_{U,1} \widehat{S} \widehat{R}_{V,1}$ for some $\widehat{Q} \in \widehat{l}_{p \times q}^1$. The free parameter \widehat{S} satisfies $\widehat{S} = \widehat{U}_0 \widehat{K} \widehat{V}_0$, if \widehat{U}_0 and \widehat{V}_0 are chosen as in (10). The operator A has been defined in such a way that it evaluates certain derivatives of components of $\widehat{S} = \widehat{U}_0 \widehat{K} \widehat{V}_0$ at certain points of D , and since \widehat{S} is a free parameter, this evaluation operator is surjective. Thus, $[A, B^*]$ is surjective, and $[A, B^*]^*$ is one-to-one.

The proof of (iv) is immediate. \square

There is a uniqueness result for the solution of (DUAL') similar to Theorem 10.2.

THEOREM 11.3. *Let (μ, Φ^+, Φ^-) be an optimal solution of (PRIM). Define $\Phi = \Phi^+ - \Phi^-$, and define the sets I and I' of indices $i \in \{1, \dots, m\}$ and the sets J , J^+ , and J^- , of indices $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$, and $k \in \{0, 1, \dots\}$ as follows:*

$$\begin{aligned} I &= \{i \mid \sum_{jk} |\Phi_{ij}(k)| = \mu\}, \\ I' &= \{i \mid \sum_{jk} |\Phi_{ij}(k)| < \mu\}, \\ J &= \{(i, j, k) \mid \Phi_{ij}(k) = 0\}, \\ J^+ &= \{(i, j, k) \mid \Phi_{ij}(k) > 0\}, \\ J^- &= \{(i, j, k) \mid \Phi_{ij}(k) < 0\}. \end{aligned}$$

Then the optimal solution (ν, τ, δ) of (DUAL) is unique if and only if the following conditions determine $\nu \in \mathbf{R}$, $\tau \in \mathbf{R}^m$, and $\eta \in l_{m \times n}^\infty$ uniquely:

$$\begin{aligned} \nu &= \langle H, \delta \rangle, \\ \sum_{i=1}^m \tau_i &= 1, \\ \delta &\in \mathcal{N}([A, B^*])^\perp, \\ \tau_i &\geq 0, \quad i \in I, \\ \tau_i &= 0, \quad i \in I', \\ -\tau_i &\leq \delta_{ij}(k) \leq \tau_i, \quad (i, j, k) \in J, \\ \delta_{ij}(k) &= \tau_i, \quad (i, j, k) \in J^+, \\ \delta_{ij}(k) &= -\tau_i, \quad (i, j, k) \in J^-. \end{aligned} \tag{36}$$

The easy proof of this theorem is left to the reader. As in the case of Theorem 10.2, conditions (36) do not refer to the primal solution directly, but only to the index sets.

12. Computation of suboptimal solutions in the FME scheme. The FMV method has one definite advantage over the FME method: at each step we have access to a suboptimal solution Φ_N of the full problem. Whenever we are ready to decide that the suboptimal value μ_N is sufficiently close to the optimal value μ , no further computations are needed. This should be contrasted to the fact that in the FME method, the functions Φ_M that we get are not feasible for the full problem, and although the FME method produces information about the optimal value μ , it does not directly produce a suboptimal solution whose norm is close to this value.

There is one possibility to bypass this problem. Recall that in the FME method, all the suboptimal solutions satisfy $A(H - \Phi_M) = 0$, i.e., $H - \Phi_M \in \mathcal{N}(A)$. By (11), with \hat{K} replaced by $\hat{H} - \hat{\Phi}_M$, this means that there is some function $Q_M \in l_{p \times q}^1$ such that $\hat{U}_0(\hat{H} - \hat{\Phi}_M)\hat{V}_0 = \hat{N}_U \hat{R}_U \hat{Q}_M \hat{L}_V \hat{N}_V$, where \hat{U}_0 and \hat{V}_0 are the matrices in (10). Solve this for \hat{Q}_M to get

$$(37) \quad \hat{Q}_M = [\hat{R}_U]^{-1}[\hat{N}_U]^{-1}\hat{U}_0(\hat{H} - \hat{\Phi}_M)\hat{V}_0[\hat{N}_V]^{-1}[\hat{L}_V]^{-1}.$$

Observe that, by Theorem 4.4, this would be the correct formula for the computation of \hat{Q} in case $\hat{\Phi}_M$ would be a feasible solution of (PRIM). Define the function Ψ_M by

$$(38) \quad \hat{\Psi}_M = \hat{H} - \hat{U}\hat{Q}_M\hat{V}.$$

Then $\hat{\Psi}_M$ can also be written in the form

$$(39) \quad \hat{\Psi}_M = \hat{H} - \hat{L}_{U,1}\hat{U}_0(\hat{H} - \hat{\Phi}_M)\hat{V}_0\hat{R}_{V,1}.$$

Clearly, by (38) this function Ψ_M is a feasible solution of (PRIM).

THEOREM 12.1. *Let $\{\Phi_{M_\ell}\}$ be a subsequence of solutions obtained from the FME scheme converging weak* to an optimal solution Φ of (PRIM), and define Ψ_{M_ℓ} as in (39), where \hat{U}_0 and \hat{V}_0 are chosen to satisfy (10). Then each Ψ_{M_ℓ} is a suboptimal solution of (PRIM), and $\Psi_{M_\ell} \rightarrow \Phi$ weak*. Moreover, if Φ_{M_ℓ} tends to Φ in the norm of $l_{m \times n}^1$, then so does Ψ_{M_ℓ} . In particular, if we define $\mu_{M_\ell} = \|\Psi_{M_\ell}\|_{l^1}$, then $\nu_{M_\ell} \leq \mu \leq \mu_{M_\ell}$, and, whenever we have norm convergence, $\mu_{M_\ell} - \nu_{M_\ell} \rightarrow 0$.*

In particular, observe that if we have norm convergence, then it is possible to get a converging upper bound on μ without the use of the FMV scheme. If we do not have norm convergence, then it may still be true that we have norm convergence for some reduced problem; see Theorems 9.1 and 11.1(ii).

Proof. The operator defined in (37) that maps $K = H - \Phi_M$ to Q_M is both norm continuous and weak* continuous from $\mathcal{N}(A)$ into $l_{p \times q}^\infty$. Thus, Q_{M_ℓ} tends to the function Q corresponding to the optimal solution Φ , either weak* or in norm, depending on the type of convergence of Φ_{N_ℓ} to Φ . By the same continuity argument, $\Psi_{M_\ell} = H - U * Q_{M_\ell} * V$ tends to $H - U * Q * V = \Phi$. The final claim is obvious. \square

13. About degeneracy. The major part of this paper deals with the two questions of how we should reformulate the model matching problem in l^1 as a linear programming problem in such a way that we avoid unnecessary redundancies, and how different formulations of the problem affect the behavior of the solutions of the two approximation schemes. In other words, we do our best to avoid degeneracies and describe how different degeneracies affect the solutions.

It is quite common to have degeneracies in finite-dimensional linear programming problems, and all standard solvers are required to cope with these in one way or another. In general, if the dimension of the problem is small, then possible degeneracies

do not cause serious trouble. If, however, the dimension of the problem is large, then degeneracies may complicate the solution process and may lead to problems that are quite hard to solve.

In our case the system is infinite-dimensional. The author has solved a number of different problems numerically, and the one thing that has caused the most trouble is the existence of a different type of degeneracies. Some of these problems caused by degeneracies are described in [14] and [15].

One way to characterize the nondegeneracy of the pair of finite-dimensional approximate problems $(\text{PRIM})_{MN}$ and $(\text{DUAL})_{MN}$ is the following. Since these problems are finite dimensional, they have optimal solutions. The pair of problems is nondegenerate if the two optimal solutions are unique, and if, in addition, the following conditions hold for all relevant values of i , j , and k (cf. Theorems 10.2 and 11.3):

$$\begin{aligned}\sum_{jk} |\Phi_{ij}(k)| &= \mu \iff \tau_i > 0, \\ \Phi_{ij}(k) > 0 &\iff (A^* \eta + B^* * \gamma)_{ij}(k) = \tau_i, \\ \Phi_{ij}(k) < 0 &\iff (A^* \eta + B^* * \gamma)_{ij}(k) = -\tau_i.\end{aligned}$$

Note that these conditions, together with the inequalities given in the original problems, imply that

$$\begin{aligned}\sum_{jk} |\Phi_{ij}(k)| < \mu &\iff \tau_i = 0, \\ \Phi_{ij}(k) = 0 &\iff -\tau_i < (A^* \eta + B^* * \gamma)_{ij}(k) < \tau_i.\end{aligned}$$

In particular (as already seen in §7), we must have $\sum_{jk} |\Phi_{ij}(k)| = \mu$ and $\tau_i > 0$ for all i .

In the infinite-dimensional case we could still use the same definition of degeneracy, provided both the primal and dual problems have optimal solutions. As we saw in Theorem 6.2, this is indeed the case whenever the range of $[A, B^*]$ is closed—a problem that we studied in Lemma 6.4. Uniqueness of the optimal dual solution requires the operator $[A, B^*]$ to have a dense range. If the range is closed, then this is equivalent to $[A, B^*]$ being surjective. A condition for this to be the case is given in Theorem 11.3(iii). The numerical problems described on pp. 179–180 of [14] are due to the fact that there the operator $[A, B^*]$ did not have this property.

The properties just described, which can be summarized as the property that $[A, B^*]$ should be surjective, is the only nondegeneracy property that it seems to be possible to build into the system without any a priori knowledge of optimal primal and dual solutions. In particular, since the property of having one or more rows i in the problem where an optimal solution Φ satisfies $\sum_{j,k} |\Phi_{ij}(k)| < \mu = \max_i \sum_{j,k} |\Phi_{ij}(k)|$ is not dependent on the particular linear programming formulation, but only on the original data, this situation cannot be avoided by a different choice of the operators A and B^* . Recall that this means massive nonuniqueness of the primal solution in the sense that in those rows the decomposition of $\Phi_{ij}(k)$ into $\Phi_{ij}(k) = \Phi_{ij}^+(k) - \Phi_{ij}^-(k)$ is no longer unique. However, in some cases this need not be a serious problem, because of the fact that although the decomposition need not be unique, the primal variable Φ itself may be unique; cf. Theorem 9.1, Lemma 9.2, and Theorem 10.1(iii).

REFERENCES

- [1] J. A. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, Birkhäuser Verlag, Basel, 1990.
- [2] J. A. BALL AND M. RAKOWSKI, *Interpolation by rational matrix functions and stability of feedback systems: the 2-block case*, J. Math. Systems and Control, to appear.
- [3] M. A. DAHLEH, *Robustness for coprime factor perturbations*, preprint, 1989.
- [4] M. A. DAHLEH AND J. B. PEARSON, l^1 -optimal feedback controllers for MIMO discrete-time systems, IEEE Trans. Automat. Control, 32 (1987), pp. 314–322.
- [5] ———, *Optimal rejection of persistent disturbances, robust stability, and mixed sensitivity minimization*, IEEE Trans. Automat. Control, 33 (1988), pp. 722–731.
- [6] B. A. FRANCIS, *A Course in H^∞ Control Theory*, Springer-Verlag, Berlin, New York, 1987.
- [7] G. S. JORDAN, O. J. STAFFANS, AND R. L. WHEELER, *Local analyticity in weighted l^1 -spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749–782.
- [8] ———, *Convolution operators in a fading memory space: the critical case*, SIAM J. Math. Anal., 18 (1987), pp. 366–386.
- [9] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [10] J. S. McDONALD AND J. B. PEARSON, l^1 -optimal control of multivariable systems with output norm constraints, Automatica, 27 (1991), pp. 317–329.
- [11] M. A. MENDLOVITZ, *A simple solution to the l^1 optimization problem*, Systems Control Lett., 12 (1989), pp. 461–463.
- [12] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [13] O. J. STAFFANS, *Sensitivity l^1 -minimization with boundary zeros and poles*, Helsinki University of Technology, Institute of Mathematics, Research Report A301, October 1991.
- [14] ———, *Mixed sensitivity minimization problems with rational l^1 -optimal solutions*, J. Optim. Theory Appl., 70 (1991), pp. 173–189.
- [15] ———, *MIMO l^1 -optimization with a scalar control*, J. Optim. Theory Appl., 74 (1992), pp. 545–564.
- [16] M. VIDYASAGAR, *Control System Synthesis: a Factorization Approach*, MIT Press, Cambridge, MA, 1985.
- [17] ———, *Further results on the optimal rejection of persistent bounded disturbances*, IEEE Trans. Automat. Control, 36 (1991), pp. 642–652.

CONTROL OF PLANAR NETWORKS OF TIMOSHENKO BEAMS*

J. E. LAGNESE[†], G. LEUGERING[‡], AND E. J. P. G. SCHMIDT[§]

Abstract. The present study is concerned with the questions of controllability and stabilizability of planar networks of vibrating beams consisting of several Timoshenko beams connected to each other by rigid joints at all interior nodes of the system. Some of the exterior nodes are either clamped or free; controls may be applied at the remaining exterior nodes and/or at interior joints in the form of forces and/or bending moments. For a given configuration, is it at all possible to drive all vibrations to the rest configuration in a given finite time interval by means of controls acting at some or all of the available (nonclamped) nodes of the network and, if so, where should such controls be placed? Alternatively, a control objective is to construct energy absorbing boundary-feedback controls that will guarantee uniform energy decay. It is demonstrated that if such a network does not contain closed loops and if at most one of the exterior nodes is clamped, exact controllability and uniform stabilizability of the network is indeed possible by means of controls placed at the free exterior nodes of the system. On the other hand, examples are presented to demonstrate that when a closed loop is present in the network or if the network has more than one clamped exterior node, it may happen that approximate control of the network to its rest configuration is not possible even if controls are placed at every available node of the system.

Key words. control of networks, Timoshenko beams, exact controllability, uniform stabilizability

AMS(MOS) subject classifications. 93C20, 93D15, 35B45

1. Introduction. In this paper are considered the questions of exact controllability and uniform stabilizability of a planar network of Timoshenko beams. Such a system is a particular case of a very general model of three-dimensional beam networks derived in [3] and to which we refer the reader for justification of the joint conditions given below.

The basic assumptions in our model are, first, that each beam of the network is adequately modeled by the Timoshenko beam system as far as transverse displacement and shear angle are concerned, and regarding longitudinal motion by the equation for the small axial deformations of a thin rod. In a system so described, the deformation of each beam is determined solely by the deformation of its centerline. The second assumption is that the deformed centerlines of the collection of beams together form a connected planar graph throughout the deformation process. This is a requirement of planar motion and also of *continuity* of the network at the interior *joints* (where two or more beams meet). The third assumption is that the interior joints are *rigid*. For the system considered here, this means that the rotation angles at a joint of each pair of beams meeting at that joint are the same throughout the deformation process; see [3] for justification of this definition. Obviously, there are other types of joint interactions that could equally well be considered and which will be the subject of subsequent studies.

The requirements of continuity and rigidity of the joints will be referred to as *geometric joint conditions*. In addition, there are *dynamic joint conditions*, repre-

* Received by the editors July 22, 1991; accepted for publication (in revised form) March 16, 1992.

[†] Department of Mathematics, Georgetown University, Washington, DC 20057. This research was supported by Air Force Office of Scientific Research grant AFOSR 88-0337.

[‡] Department of Mathematics, Georgetown University, Washington, DC 20057. This research was supported by the Deutsche Forschungsgemeinschaft (DFG), Heisenbergreferat, L-595-3-1.

[§] Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6.

senting balances of forces and moments at each joint. These may be obtained by direct analysis of the forces and moments acting on a joint or, more simply, from the variational formulation of the boundary value problem associated with the beam network, keeping in mind that the variation must be taken with respect to displacements that satisfy the geometric joint conditions described above, in addition to the usual geometric conditions at “simple nodes” (i.e., a node where only one beam begins or ends). Controls are introduced into the structure at the free exterior nodes through appropriate forces and moments and/or at the interior joints through the dynamic joint conditions. The control objective is to bring the entire system to its equilibrium configuration in a finite time, independent of the initial state of the system (exact null controllability). Alternatively, through the application of appropriate energy absorbing type feedback controls, we seek to damp out the vibrations in the structure at a uniform rate (uniform stabilization). We will find that structures for which at most one exterior node is clamped and which do not contain closed loops are both exactly controllable and uniformly stabilizable using controls at only the free exterior nodes of the network. On the other hand, examples will be given that demonstrate that in situations where either of these two conditions is violated, there may exist periodic waves propagating within the structure that cannot be influenced by means of nodal controls; this will be seen to be a highly “nongeneric” situation, however.

To the authors’ knowledge, to date very little research has been done on the problem of controllability, or stabilizability, of multiple-link structures from the point of view of distributed parameter systems. Closest to the spirit of the present paper is the work of Schmidt [9] dealing with exact controllability of a network of vibrating strings, and that of Leugering and Schmidt [6], where exact controllability of a network of rigidly connected Euler–Bernoulli beams was considered using so-called “slide and turn” controls (controls acting in the geometric boundary conditions, rather than through forces and moments, at exterior nodes). However, in [6] exact controllability was established only for a “bundle” of beams, i.e., a collection of beams, all of which emanate from a single multiple joint, with controls applied at all of the exterior nodes. The controllability results of the present paper are much more general in the sense that they apply to a richer class of beam configurations. We should also mention the paper of Chen et al., [1], where the issue of uniform stabilization of a *serially connected* network of Euler–Bernoulli beams was studied, and the work of Lions on exact controllability of “problems of transmission”; see [7, Chap. VI].

2. Description of the model. The network is comprised of n Timoshenko beams, labeled with indices $[1, 2, \dots, n]$. The equilibrium state of the centerline of the i th beam is a straight line segment of length ℓ_i , parametrized by x , $0 \leq x \leq \ell_i$, and is described parametrically by

$$\mathbf{p}_{0i} + x\mathbf{e}_i, \quad 0 \leq x \leq \ell_i,$$

where \mathbf{p}_{0i} is a fixed position vector in \mathbb{R}^3 and \mathbf{e}_i a unit vector in \mathbb{R}^3 directed along the centerline of the i th beam. The \mathbf{e}_i ’s are assumed to be coplanar, and the collection of reference lines is supposed to form a connected set. (We have tacitly assumed that the beam is not subject to tensile loading.) Let \mathbf{e}_i^\perp be a unit vector in the plane containing all of the direction vectors that is orthogonal to \mathbf{e}_i . In Timoshenko beam theory the vector pointing to the deformed material point originally located along the filament

$$\mathbf{p}_{0i} + x\mathbf{e}_i + z\mathbf{e}_i^\perp, \quad |z| \leq h/2,$$

is given by

$$\begin{aligned}\mathbf{R}_i(x, z, t) &:= \mathbf{p}_{0i} + (x + u_i(x, t) + z\psi_i(x, t))\mathbf{e}_i + (z + w_i(x, t))\mathbf{e}_i^\perp \\ &= (\mathbf{p}_{0i} + x\mathbf{e}_i + z\mathbf{e}_i^\perp) + \mathbf{r}_i(x, t) + z\psi_i(x, t)\mathbf{e}_i, \quad i = 1, \dots, n,\end{aligned}$$

where $\mathbf{r}_i = u_i\mathbf{e}_i + w_i\mathbf{e}_i^\perp$, u_i, w_i denote longitudinal and transverse displacements, respectively, of the point originally situated at $\mathbf{p}_{0i} + x\mathbf{e}_i$, and ψ_i is the rotation angle of the filament ($\phi_i = \psi_i + w'_i$ is the *shear angle*).

With this notation, the dynamic equations of motion of the i th beam (in the absence of body forces) are

$$\begin{aligned}(2.1) \quad & \rho_i \ddot{u}_i - E_i A_i u''_i = 0, \\ & \rho_i \ddot{w}_i - K_i (\psi_i + w'_i)' = 0, \\ & I_{\rho_i} \ddot{\psi}_i - E_i I_i \psi''_i + K_i (\psi_i + w'_i) = 0, \quad 0 < x < \ell_i, \quad t > 0.\end{aligned}$$

We have used $\dot{}$ and $'$ to denote time and spatial differentiation, respectively. The physical constants appearing in the above system are ρ_i , the mass density per unit of reference length; A_i , the area of a cross section in the reference configuration; E_i , Young's modulus of elasticity; I_i , the second moment of inertia of a cross section; I_{ρ_i} , the polar moment of inertia of a cross section; and K_i , the shear modulus. Although u_i is uncoupled from the other dependent variables in the above system, it is coupled to them through joint conditions, which we now describe.

A *node* of the beam network is a point in the plane where a beam begins or ends, i.e., an endpoint of a centerline. (We identify a beam with its reference line.) The nodes consist of *simple nodes*, where only one beam begins or ends, and *multiple nodes* (or *joints*), where two or more beams join. Suppose there are a total of m nodes, and label them with indices $[1, 2, \dots, m] := I$. We set $I = I_S \cup I_M$, where

$$\begin{aligned}I_S &= \{i \in I \mid i \text{ belongs to a simple node}\}, \\ I_M &= \{i \in I \mid i \text{ belongs to a multiple node}\}.\end{aligned}$$

We further partition $I_S = I_S^D \cup I_S^N$, $I_M = I_M^D \cup I_M^N$, where

$$\begin{aligned}I_S^D &= \{i \in I_S \mid i \text{ belongs to a clamped simple node}\}, \\ I_S^N &= \{i \in I_S \mid i \text{ belongs to a free simple node}\}, \\ I_M^D &= \{i \in I_S \mid i \text{ belongs to a clamped multiple node}\}, \\ I_M^N &= \{i \in I_S \mid i \text{ belongs to a free multiple node}\}.\end{aligned}$$

(Here we are tacitly assuming that each node belongs to one of these four classes.) To simplify the notation slightly, if $I_S^N \neq \emptyset$ we assume that the beams and nodes have been labeled so that $I_S^N = [1, \dots, p]$ for some $p < m$, and that the beam ending at node N_i ($i=1, \dots, p$) is beam i .

To describe the boundary conditions, it is convenient to introduce the following notation. Let N_k be a node corresponding to an index $k \in I$. Following Schmidt [9] we set

$$\mathcal{E}_k = \{i \in [1, 2, \dots, n] \mid N_k \text{ is an endpoint of beam } i\}.$$

Thus \mathcal{E}_k is a singleton if N_k is a simple node. For each $i \in \mathcal{E}_k$ we set

$$\varepsilon_{ik} = \begin{cases} -1 & \text{if } N_k \text{ corresponds to } x = 0, \\ +1 & \text{if } N_k \text{ corresponds to } x = \ell_i. \end{cases}$$

Then the outward pointing spatial derivative at the i th beam and at node N_k may be written $\varepsilon_{ik}(\partial/\partial x)$. The set of signs of the ε_{ik} at a multiple node is called the *sign arrangement* at N_k . We adopt the convention that

$$\varepsilon_{ik} = \begin{cases} -1 & \text{if } N_k \text{ is associated to an index } i \in I_S^D, \\ +1 & \text{if } N_k \text{ is associated to an index } i \in I_S^N. \end{cases}$$

Thus $x = 0$ at each clamped simple node, while $x = \ell_i$ at a free simple node.

Set $I^D = I_S^D \cup I_M^D$, which contains the indices of the clamped nodes. At such a node the boundary conditions are

$$(2.2) \quad u_i(N_k, t) = w_i(N_k, t) = \psi_i(N_k, t) = 0, \quad \forall i \in \mathcal{E}_k, \quad k \in I^D.$$

At a free simple node the boundary conditions are

$$(2.3) \quad E_i A_i u'_i(N_i, t) = K_i(\psi_i + w'_i)(N_i, t) = E_i I_i \psi'_i(N_i, t) = 0, \quad i = 1, \dots, p.$$

The geometric and dynamic conditions at a free multiple mode are, respectively,

$$(2.4) \quad \mathbf{r}_i(N_k, t) = \mathbf{r}_j(N_k, t), \quad \psi_i(N_k, t) = \psi_j(N_k, t), \quad \forall i, j \in \mathcal{E}_k, \quad \forall k \in I_M^N,$$

and

$$\begin{cases} \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} E_i I_i \psi'_i(N_k, t) = 0, \\ \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} [E_i A_i u'_i \mathbf{e}_i + K_i(\psi_i + w'_i) \mathbf{e}_i^\perp](N_k, t) = 0, \quad \forall k \in I_M^N. \end{cases}$$

Introducing the vector $\mathbf{n} := \mathbf{e}_i \times \mathbf{e}_i^\perp$ (which is independent of i) allows the last set of equations to be written

$$(2.5) \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} [E_i A_i u'_i \mathbf{e}_i + K_i(\psi_i + w'_i) \mathbf{e}_i^\perp + E_i I_i \psi'_i \mathbf{n}](N_k, t) = 0, \quad \forall k \in I_M^N.$$

The dynamic joint conditions in their present forms assert that the resultant of bending moments and resultant of forces at each free multiple joint are zero. There is a slight abuse of notation in (2.4), (2.5); for instance, $\mathbf{r}_i(N_k, t)$ stands for $\mathbf{r}_i(0, t)$ or $\mathbf{r}_i(\ell_i, t)$, depending on whether N_k corresponds to $x = 0$ or to $x = \ell_i$.

The dynamic description of the network is completed by prescribing its initial state:

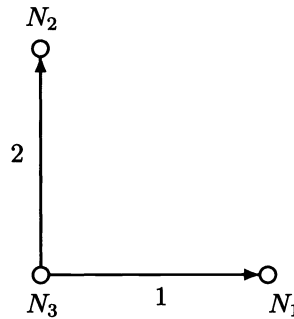
$$(2.6) \quad \mathbf{r}_i(0) = \mathbf{r}_i^0, \quad \dot{\mathbf{r}}_i(0) = \mathbf{r}_i^1, \quad \psi_i(0) = \psi_i^0, \quad \dot{\psi}_i(0) = \psi_i^1, \quad i = 1, 2, \dots, n,$$

with the usual notational convention: $\mathbf{r}_i(0)$ stands for $\mathbf{r}_i(x, 0)$, $0 \leq x \leq \ell_i$, etc.

Example 2.1. Consider the case of the “carpenter’s square” in \mathbb{R}^2 that is free at each simple node, as shown in Fig. 1.

Here we may choose $\mathbf{e}_1 = (1, 0)$, $\mathbf{e}_2 = (0, 1)$, so that $\mathbf{e}_1^\perp = \mathbf{e}_2$ and $\mathbf{e}_2^\perp = -\mathbf{e}_1$. The geometric node conditions at N_1 are therefore

$$u_1(0, t) = -w_2(0, t), \quad u_2(0, t) = w_1(0, t), \quad \psi_1(0, t) = \psi_2(0, t),$$

FIG. 1. *Carpenter's square configuration.*

while the dynamic node conditions there are

$$\begin{aligned} E_1 A_1 u'_1(0, t) - K_2(\psi_2 + w'_2)(0, t) &= 0, \\ E_2 A_2 u'_2(0, t) + K_1(\psi_1 + w'_1)(0, t) &= 0, \\ E_1 I_1 \psi'_1(0, t) + E_2 I_2 \psi'_2(0, t) &= 0. \end{aligned}$$

If N_2 is clamped rather than free, then $\mathbf{e}_2 = (0, -1)$, $\mathbf{e}_1^\perp = -\mathbf{e}_2$ and $\mathbf{e}_2^\perp = \mathbf{e}_1$, so that the geometric and dynamic node conditions then become

$$u_1(0, t) = w_2(\ell_2, t), \quad u_2(\ell_2, t) = -w_1(0, t), \quad \psi_1(0, t) = \psi_2(\ell_2, t),$$

and

$$\begin{aligned} E_1 A_1 u'_1(0, t) - K_2(\psi_2 + w'_2)(\ell_2, t) &= 0, \\ E_2 A_2 u'_2(\ell_2, t) + K_1(\psi_1 + w'_1)(0, t) &= 0, \\ E_1 I_1 \psi'_1(0, t) - E_2 I_2 \psi'_2(\ell_2, t) &= 0, \end{aligned}$$

respectively.

3. Controllability estimates for the Timoshenko network. In this section the a priori estimates that are needed in the study of exact controllability and uniform stabilizability of the Timoshenko network will be established.

The *total energy* of the network is given by

$$\mathcal{Q}(t) = \frac{1}{2} \sum_{i=1}^n \left\{ \int_0^{\ell_i} [\rho_i |\dot{\mathbf{r}}_i|^2 + I_{\rho_i} \dot{\psi}_i^2] dx + \int_0^{\ell_i} [E_i A_i u_i'^2 + E_i I_i \psi_i'^2 + K_i (\psi_i + w_i')^2] dx \right\}.$$

The sum of the first integrals represents the kinetic energy and the sum of the second the strain energy of the network. Denote by $Q_i(x, t)$ the energy densities:

$$\mathcal{Q}(t) = \frac{1}{2} \sum_{i=1}^n \int_0^{\ell_i} Q_i(x, t) dx.$$

If u, w, ψ is a sufficiently regular solution of (2.1), the time rate of change of the corresponding energy functional is given by

$$(3.1) \quad \dot{\mathcal{Q}}(t) = \frac{1}{2} \sum_{i=1}^n \int_0^{\ell_i} \dot{Q}_i(x, t) dx$$

$$= \sum_{i=1}^n [E_i A_i u'_i \dot{u}_i + E_i I_i \psi'_i \dot{\psi}_i + K_i (\psi_i + w'_i) \dot{w}_i] \Big|_0^{\ell_i}.$$

At each free simple node and at each clamped node the corresponding terms in the right-hand sum vanish. The rest of the sum may be written

$$(3.2) \quad \sum_{k \in I_M^N} \left(\sum_{i \in \mathcal{E}_k} \varepsilon_{ik} [E_i A_i u'_i \mathbf{e}_i + K_i (\psi_i + w'_i) \mathbf{e}_i^\perp] (N_k, t) \right) \cdot \dot{\mathbf{r}}_{i_k} \\ + \sum_{k \in I_M^N} \left(\sum_{i \in \mathcal{E}_k} \varepsilon_{ik} E_i I_i \psi'_i (N_k, t) \right) \dot{\psi}_{i_k} = 0$$

in view of (2.4) and (2.5), where the index i_k in (3.2) stands for any fixed index in \mathcal{E}_k , for example, $i_k = \min\{i \in \mathcal{E}_k\}$. Therefore, *the total energy of the system (2.1)–(2.5) is time invariant.*

Before proceeding to the derivation of energy estimates, we must say a little bit about existence and regularity of solutions of (2.1)–(2.6). We set

$$H = \prod_{i=1}^n L^2((0, \ell_i); \mathbb{R}^3) \\ = \prod_{i=1}^n L^2(0, \ell_i; [\mathbf{e}_i]) \oplus L^2(0, \ell_i; [\mathbf{e}_i^\perp]) \oplus L^2(0, \ell_i; [\mathbf{n}]),$$

where $[\mathbf{e}]$ denotes the linear span of a unit vector \mathbf{e} . If $\mathbf{R} \in H$ we write

$$\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_n), \quad \mathbf{R}_i = u_i \mathbf{e}_i + w_i \mathbf{e}_i^\perp + \psi_i \mathbf{n} = \mathbf{r}_i + \psi_i \mathbf{n}.$$

The norm on H is taken to be

$$\|\mathbf{R}\|_H = \left[\sum_{i=1}^n \int_0^{\ell_i} (\rho_i |\mathbf{r}_i|^2 + I_{\rho_i} \psi_i^2) dx \right]^{1/2}.$$

We also define

$$V = \left\{ \mathbf{R} \in \prod_{i=1}^n H^1((0, \ell_i); \mathbb{R}^3) \mid \mathbf{R} \text{ satisfies (2.2) and (2.4)} \right\}.$$

If $I^D \neq \emptyset$, the following defines a norm on V equivalent to the norm induced on V by $\prod_{i=1}^n H^1((0, \ell_i); \mathbb{R}^3)$ with its standard topology:

$$\|\mathbf{R}\|_V = \left\{ \sum_{i=1}^n \int_0^{\ell_i} [E_i A_i u_i'^2 + E_i I_i \psi_i'^2 + K_i (\psi_i + w_i')^2] dx \right\}^{1/2}.$$

If $I^D = \emptyset$, V may be defined as a quotient space modulo the rigid motions $\mathbf{R}_i = a_i \mathbf{e}_i + (c_i - b_i x) \mathbf{e}_i^\perp + b_i \mathbf{n}$ that satisfy (2.4). We note that

$$\mathcal{Q}(t) = \frac{1}{2} \left(\|\dot{\mathbf{R}}(t)\|_H^2 + \|\mathbf{R}(t)\|_V^2 \right).$$

The system (2.1)–(2.5) may be written as a variational equation as follows. Let \mathbf{R} be a solution of (2.1)–(2.5) and $\hat{\mathbf{R}} \in V$ be a test function. Form the $L^2((0, \ell_i); \mathbb{R}^3)$ scalar product of (2.1) with $\hat{\mathbf{R}}_i$ and sum over i from 1 to n . We obtain, with the aid of (2.2)–(2.5), the variational equation

$$(\ddot{\mathbf{R}}, \hat{\mathbf{R}})_H + (\mathbf{R}, \hat{\mathbf{R}})_V = 0, \quad \forall \hat{\mathbf{R}} \in V$$

or, equivalently,

$$(3.3) \quad \ddot{\mathbf{R}} + A\mathbf{R} = 0 \quad \text{in } V',$$

where V' denotes the dual space of V with respect to H and A is the Riesz isomorphism of V onto V' . The initial conditions (2.6) are written

$$(3.4) \quad \mathbf{R}(0) = \mathbf{R}^0, \quad \dot{\mathbf{R}}(0) = \mathbf{R}^1.$$

It follows from standard theory that if $(\mathbf{R}^0, \mathbf{R}^1) \in H \times V'$, then (3.3), (3.4) has a unique solution with $(\mathbf{R}, \dot{\mathbf{R}}) \in C([0, \infty); H \times V')$, called a *weak solution*; if $(\mathbf{R}^0, \mathbf{R}^1) \in V \times H$, then $(\mathbf{R}, \dot{\mathbf{R}})$ has the regularity $C([0, \infty); V \times H)$, and we then call the solution a *finite energy solution*; and if $(\mathbf{R}^0, \mathbf{R}^1) \in D_A \times V$ then $(\mathbf{R}, \dot{\mathbf{R}})$ belongs to $C([0, \infty); D_A \times V)$ and the solution called a *classical solution*, where

$$D_A = \{\mathbf{R} \in V \mid A\mathbf{R} \in H\}, \quad \|\mathbf{R}\|_{D_A} = \|A\mathbf{R}\|_H.$$

It is not difficult to verify that

$$\mathbf{R} \in D_A \iff \mathbf{R}_i \in H^2((0, \ell_i); \mathbb{R}^3) \text{ and satisfies (2.2)–(2.5), } i = 1, \dots, n.$$

An energy identity that is the basis for the a priori estimates to follow will now be derived.

PROPOSITION 3.1. *The following identity holds for each $T > 0$ and each classical solution of (2.1):*

$$(3.5) \quad f(T) - f(0) + \frac{1}{2} \sum_{i=1}^n \int_0^T \int_0^{\ell_i} [\alpha'_i Q_i - 2\alpha_i \rho_i \dot{w}_i \dot{\psi}_i + 2\alpha_i K_i \psi'_i (\psi_i + w'_i)] dx dt \\ = \frac{1}{2} \sum_{k=1}^m \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) \int_0^T Q_i(N_k, t) dt,$$

where α_i is any $C^1([0, \ell_i])$ function and where

$$f(t) = \sum_{i=1}^n \int_0^{\ell_i} \alpha_i [\rho_i \dot{u}_i u'_i + \rho_i \dot{w}_i (\psi_i + w'_i) + I_{\rho_i} \dot{\psi}_i \psi'_i] dx.$$

A corollary of (3.5) is the following energy estimate.

COROLLARY 3.2. *Let*

$$k_i = \max \left(\sqrt{\frac{\rho_i}{I_{\rho_i}}}, \sqrt{\frac{K_i}{E_i I_i}} \right),$$

and let α_i be a positive, strictly increasing C^1 function on $[0, \ell_i]$ such that $\alpha'_i/\alpha_i > k_i$, $i = 1, 2, \dots, n$. For each $T > 0$ and each finite energy solution of (2.1)–(2.5) there are positive constants c_0, T_0 such that

$$(3.6) \quad c_0(T - T_0)\mathcal{Q}(0) \leq \frac{1}{2} \sum_{i=1}^p \alpha_i(N_i) \int_0^T [\rho_i |\dot{\mathbf{r}}_i|^2 + I_{\rho_i} \dot{\psi}_i^2](N_i, t) dt \\ + \frac{1}{2} \sum_{k \in I_M} \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) \int_0^T Q_i(N_k, t) dt.$$

Proof of Proposition 3.1. Identity (3.5) is proved by the use of multipliers and integrations by parts in the standard way. We multiply the three equations in (2.1) by $\alpha_i u'_i$, $\alpha_i w'_i$, and $\alpha_i \psi'_i$, respectively, add the three products, and apply $\int_0^T \int_0^{\ell_i}$ to the sum. (We might, of course, try different multipliers for each equation or even multiply (2.1) by $\mathcal{M}_i(x)(u'_i \ w'_i \ \psi'_i)^*$ using a 3×3 matrix \mathcal{M}_i of multipliers. But to obtain something useful we are forced in the end to choose \mathcal{M}_i diagonal and, moreover, nothing extra is gained by choosing distinct diagonal elements.) We obtain after an integration by parts

$$(3.7) \quad f_{1i}(T) - f_{1i}(0) + \frac{1}{2} \int_0^T \int_0^{\ell_i} \alpha'_i [\rho_i (\dot{u}_i^2 + \dot{w}_i^2) + I_{\rho_i} \dot{\psi}_i^2] dx dt \\ - \frac{1}{2} \int_0^T \alpha_i [\rho_i (\dot{u}_i^2 + \dot{w}_i^2) + I_{\rho_i} \dot{\psi}_i^2] \Big|_0^{\ell_i} dt \\ - \int_0^T \int_0^{\ell_i} \alpha_i [E_i A_i u''_i u'_i + K_i (\psi_i + w'_i)' w'_i \\ + E_i I_i \psi''_i \psi'_i - K_i (\psi_i + w'_i) \psi'_i] dx dt = 0,$$

where

$$(3.8) \quad f_{1i}(t) = \int_0^{\ell_i} \alpha_i [\rho_i (\dot{u}_i u'_i + \dot{w}_i w'_i) + I_{\rho_i} \dot{\psi}_i \psi'_i] dx.$$

We have

$$(3.9) \quad - \int_0^T \int_0^{\ell_i} \alpha_i [E_i A_i u''_i u'_i + K_i (\psi_i + w'_i)' w'_i + E_i I_i \psi''_i \psi'_i - K_i (\psi_i + w'_i) \psi'_i] dx dt \\ = - \frac{1}{2} \int_0^T [\alpha_i (E_i A_i u_i'^2 + E_i I_i \psi_i'^2 + K_i (\psi_i + w'_i)^2)] \Big|_0^{\ell_i} dt \\ + \frac{1}{2} \int_0^T \int_0^{\ell_i} \alpha'_i [E_i A_i u_i'^2 + E_i I_i \psi_i'^2 + K_i (\psi_i + w'_i)^2] dx dt \\ + \int_0^T \int_0^{\ell_i} \alpha_i K_i [\psi_i (\psi_i + w'_i)' + (\psi_i + w'_i) \psi'_i] dx dt.$$

Substitution of (3.9) into (3.7) yields

$$(3.10) \quad f_{1i}(T) - f_{1i}(0) + \int_0^T \int_0^{\ell_i} \alpha_i K_i [\psi_i (\psi_i + w'_i)' + (\psi_i + w'_i) \psi'_i] dx dt \\ + \frac{1}{2} \int_0^T \int_0^{\ell_i} \alpha'_i(x) Q_i(x, t) dx dt - \frac{1}{2} \int_0^T \alpha_i(x) Q_i(x, t) \Big|_{x=0}^{\ell_i} dt = 0.$$

From (2.1) we have

$$(3.11) \quad \int_0^T \int_0^{\ell_i} \alpha_i K_i \psi_i (\psi_i + w'_i)' dx dt = \int_0^T \int_0^{\ell_i} \alpha_i \rho_i \psi_i \ddot{w}_i dx dt \\ = f_{2i}(T) - f_{2i}(0) - \int_0^T \int_0^{\ell_i} \alpha_i \rho_i \dot{\psi}_i \dot{w}_i dx dt,$$

where

$$f_{2i}(t) = \int_0^{\ell_i} \alpha_i \rho_i \psi_i \dot{w}_i dx.$$

It follows from (3.10) and (3.11) that

$$(3.12) \quad f_i(T) - f_i(0) + \frac{1}{2} \int_0^T \int_0^{\ell_i} [\alpha'_i Q_i - 2\alpha_i \rho_i \dot{w}_i \dot{\psi}_i + 2\alpha_i K_i \psi'_i (\psi_i + w'_i)] dx dt \\ = \frac{1}{2} \int_0^T \alpha_i Q_i \Big|_{x=0}^{\ell_i} dt,$$

where $f_i = f_{1i} + f_{2i}$. Proposition 3.1 now follows upon summing (3.12) over i from 1 to n . \square

Proof of Corollary 3.2. We have

$$\alpha'_i Q_i - 2\alpha_i \rho_i \dot{w}_i \dot{\psi}_i + 2\alpha_i K_i \psi'_i (\psi_i + w'_i) \geq c_0 Q_i$$

for some $c_0 > 0$ if and only if $\alpha'_i > 0$ and

$$(\alpha_i \rho_i)^2 - \alpha_i'^2 \rho_i I_{\rho_i} < 0 \quad \text{and} \quad (\alpha_i K_i)^2 - \alpha_i'^2 K_i (E_i I_i) < 0$$

for $0 \leq x \leq \ell_i$. If also $\alpha_i > 0$, these are the same as

$$\frac{\alpha'_i}{\alpha_i} > \sqrt{\frac{\rho_i}{I_{\rho_i}}} \quad \text{and} \quad \frac{\alpha'_i}{\alpha_i} > \sqrt{\frac{K_i}{E_i I_i}}.$$

Also, since the total energy is time invariant for each solution of (2.1)–(2.5),

$$\sum_{i=1}^n \int_0^T \int_0^{\ell_i} Q_i dx dt = T \mathcal{Q}(0),$$

and

$$|f(t)| \leq C \mathcal{Q}(0)$$

for some $C > 0$ depending on the elastic parameters and on $\max_i \max_x \alpha_i(x)$. If we further use the boundary conditions (2.2), (2.3) in the right-hand side of (3.5) we obtain (3.6) with $T_0 = 2C/c_0$. This proves the corollary for classical solutions. On the other hand, for finite energy solutions of (2.1)–(2.5) we always have the inequality

$$(3.13) \quad \sum_{k=1}^m \sum_{i \in \mathcal{E}_k} \int_0^T Q_i(N_k, t) dt \leq C(T+1) \mathcal{Q}(0)$$

for some constant C . Indeed, in (3.5) it is only necessary to choose α_i so that $\varepsilon_{ik}\alpha_i(N_k) > 0$, since the left member of (3.5) is bounded above in absolute value by $C(T+1)\mathcal{Q}(0)$, which proves (3.13) for classical solutions and then, by approximation, for finite energy solutions. Thus the right-hand side of (3.6) is finite for all finite energy solutions of (2.1)–(2.5). \square

Remark 3.1. Inequality (3.6) may be written

$$(3.14) \quad c_0(T - T_0)(\|\mathbf{R}^0\|_V^2 + \|\mathbf{R}^1\|_H^2) \leq \sum_{i=1}^p \alpha_i(N_i) \int_0^T [\rho_i |\dot{\mathbf{r}}_i|^2 + I_{\rho_i} \dot{\psi}_i^2](N_i, t) dt \\ + \sum_{k \in I_M} \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) \int_0^T Q_i(N_k, t) dt,$$

where $(\mathbf{R}^0, \mathbf{R}^1) \in V \times H$ is the initial data of the solution. Suppose now that \mathbf{R} is only a weak solution, so that $(\mathbf{R}^0, \mathbf{R}^1) \in H \times V'$, and define $\tilde{\mathbf{R}} = -A^{-1}\dot{\mathbf{R}}$. Then $\dot{\tilde{\mathbf{R}}} = \mathbf{R}$ and $\tilde{\mathbf{R}} + A\tilde{\mathbf{R}} = 0$, so that $\tilde{\mathbf{R}}$ is a finite energy solution of (3.3) with initial data $(\tilde{\mathbf{R}}(0), \dot{\tilde{\mathbf{R}}}(0)) = (-A^{-1}\mathbf{R}^1, \mathbf{R}^0)$. Upon applying (3.14) to $\tilde{\mathbf{R}}$ we obtain the estimate

$$(3.15) \quad c_0(T - T_0)(\|\mathbf{R}^0\|_H^2 + \|\mathbf{R}^1\|_{V'}^2) \leq \sum_{i=1}^p \alpha_i(N_i) \int_0^T [\rho_i |\mathbf{r}_i|^2 + I_{\rho_i} \psi_i^2](N_i, t) dt \\ + \sum_{k \in I_M} \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) \int_0^T \tilde{Q}_i(N_k, t) dt,$$

where \tilde{Q} denotes the energy density of the solution $\tilde{\mathbf{R}}$. Likewise, corresponding to (3.13) we have

$$\sum_{k=1}^m \sum_{i \in \mathcal{E}_k} \int_0^T \tilde{Q}_i(N_k, t) dt \leq C(T+1)(\|\mathbf{R}^0\|_H^2 + \|\mathbf{R}^1\|_{V'}^2),$$

from which follows, in particular, that

$$(3.16) \quad \sum_{i=1}^p \int_0^T [\rho_i |\mathbf{r}_i|^2 + I_{\rho_i} \psi_i^2](N_i, t) dt \leq C(T+1)(\|\mathbf{R}^0\|_H^2 + \|\mathbf{R}^1\|_{V'}^2).$$

3.1. Estimating the multiple node terms. We now wish to investigate under what conditions the energy estimate (3.6) can be strengthened to

$$(3.17) \quad c_0(T - T_0)(\|\mathbf{R}^0\|_V^2 + \|\mathbf{R}^1\|_H^2) \leq \sum_{i=1}^p \alpha_i(N_i) \int_0^T [\rho_i |\dot{\mathbf{r}}_i|^2 + I_{\rho_i} \dot{\psi}_i^2](N_i, t) dt.$$

Inequality (3.17) is the type of controllability estimate that is needed in situations where controls are constrained to act at the free simple nodes only. To obtain it we will investigate under what conditions on the sign arrangements ε_{ik} and on the multipliers $\alpha_i(N_k)$ it is true that the multiple node conditions imply the inequality

$$(3.18) \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) Q_i(N_k, t) \leq 0$$

at each multiple node.

It is clear that (3.18) can hold (with positive multipliers) at a clamped multiple node, i.e., for indices $k \in I_M^D$, if and only if $\varepsilon_{ik} = -1$ for all $i \in \mathcal{E}_k$. While the same sign arrangement is obviously sufficient to assure that (3.18) holds also at a free multiple node, there are other sign arrangements for which (3.18) may be verified with appropriately chosen positive values of $\alpha_i(N_k)$. While it should not be surprising that given an arbitrary sign arrangement at N_k it is not, in general, possible to choose positive values of $\alpha_i(N_k)$ such that (3.18) holds at a free multiple node, we will nevertheless find interesting configurations for which (3.18) can be established.

Writing things out, we have

$$(3.19) \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) Q_i(N_k, t) = \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) (\rho_i |\dot{\mathbf{r}}_i|^2 + I_{\rho_i} \dot{\psi}_i^2)(N_k, t) \\ + \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) [E_i A_i u_i'^2 + E_i I_i \psi_i'^2 + K_i (\psi_i + w_i')^2](N_k, t).$$

The geometric joint conditions imply that

$$(3.20) \quad \dot{\psi}_i^2(N_k, t) = \dot{\psi}_j^2(N_k, t), \quad |\dot{\mathbf{r}}_i(N_k, t)|^2 = |\dot{\mathbf{r}}_j(N_k, t)|^2, \quad \forall i, j \in \mathcal{E}_k.$$

Therefore

$$(3.21) \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) [\rho_i |\dot{\mathbf{r}}_i|^2 + I_{\rho_i} \dot{\psi}_i^2](N_k, t) \\ = \left(\sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) \rho_i \right) |\dot{\mathbf{r}}_{i_k}(N_k, t)|^2 + \left(\sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) I_{\rho_i} \right) \dot{\psi}_{i_k}^2(N_k, t),$$

where the index i_k stands for any index in \mathcal{E}_k , in accordance with the conditions (3.20). Therefore (3.21) is nonpositive if and only if

$$(3.22) \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) \rho_i \leq 0 \quad \text{and} \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) I_{\rho_i} \leq 0.$$

In particular, if the sign arrangement at N_k is $(+ \ - \ - \ \dots \ -)$, i.e., if $\varepsilon_{qk} = +1$ and $\varepsilon_{ik} = -1$ for some $q \in \mathcal{E}_k$ and all $i \neq q$ in \mathcal{E}_k , then (3.22) will be satisfied if every $\alpha_j(N_k) > 0$ and if $\alpha_q(N_k)$ is sufficiently small relative to the other values $\alpha_i(N_k)$, $i \neq q$. Lemma 3.3 below demonstrates that the same relative choices of nodal values of the multipliers will also make the second sum on the right-hand side of (3.19) nonpositive when the linear constraints (2.5) are imposed.

Let

$$X_i = \sqrt{E_i A_i} u_i' \mathbf{e}_i + \sqrt{K_i} (\psi_i + w_i') \mathbf{e}_i^\perp + \sqrt{E_i I_i} \psi_i' \mathbf{n}, \\ Y_i = E_i A_i u_i' \mathbf{e}_i + K_i (\psi_i + w_i') \mathbf{e}_i^\perp + E_i I_i \psi_i' \mathbf{n}.$$

LEMMA 3.3. *Suppose that $\varepsilon_{qk} = +1$ and $\varepsilon_{ik} = -1$ for some $q \in \mathcal{E}_k$ and all $i \neq q$ in \mathcal{E}_k . If the $\alpha_j(N_k) > 0$ are chosen so that $\alpha_q(N_k)$ is sufficiently small relative to the values of $\alpha_i(N_k)$, $i \neq q$, then*

$$\sum_{i \in \mathcal{E}_k} \varepsilon_{ik} Y_i = 0 \implies \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) |X_i|^2 \leq 0.$$

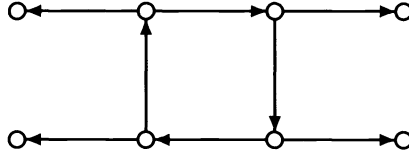


FIG. 2. Although the sign arrangements are in accord with Corollary 3.4, inequality (3.18) cannot hold simultaneously at every multiple node if the multipliers are required to be increasing functions.

Proof. There are positive constants c, C such that

$$c|Y_i|^2 \leq |X_i|^2 \leq C|Y_i|^2, \quad \forall i \in \mathcal{E}_k.$$

We write $\sum_{i \neq q}$ for the sum over indices $i \in \mathcal{E}_k$ with $i \neq q$. We have

$$\begin{aligned} \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} \alpha_i(N_k) |X_i|^2 &= \alpha_q(N_k) |X_q|^2 - \sum_{i \neq q} \alpha_i(N_k) |X_i|^2 \\ &\leq C \alpha_q(N_k) |Y_q|^2 - c \sum_{i \neq q} \alpha_i(N_k) |Y_i|^2 \\ &\leq C \alpha_q(N_k) \sum_{i \neq q} |Y_i|^2 - c \sum_{i \neq q} \alpha_i(N_k) |Y_i|^2. \end{aligned}$$

The last right-hand side will be nonpositive if α_q is chosen sufficiently small relative to the other α_i 's. \square

As a consequence of Lemma 3.3 and the discussion preceeding it, we have the following result.

COROLLARY 3.4. *If N_k is a clamped multiple node with sign arrangement $(- \dots -)$, or if N_k is a free multiple node with a sign arrangement of either $(- \dots -)$ or $(+ \dots -)$, then positive values of $\alpha_i(N_k)$ may be chosen so that the estimate (3.18) will hold.*

It should be recalled that the parametrization of the network is not completely arbitrary, but rather is constrained by sign restrictions on the ε_{ik} 's at the simple nodes.

Remark 3.2. It is not difficult to show that the sign arrangements delineated in the last corollary are, in fact, *necessary* for the validity of (3.18) (with positive multipliers) at each multiple node. In addition, if a beam network contains a *closed loop*, it is not possible to choose positive, *increasing*, multipliers for the beams in the loop (as is required in Corollary 3.2) in such a way that (3.18) holds at each joint in the loop, even when the sign arrangements are selected in accordance with Corollary 3.4 (see Fig. 2). This is because the nodal values of the multipliers must increase in the directions of the vectors \mathbf{e}_i emanating from a particular node. In problems of control of beam networks, the failure of (3.18) to hold at a nonserial joint indicates that it is necessary to introduce controls at that joint if we hope to control the motion of the entire network. This issue will be discussed further in the next section.

4. Controllability of beam networks. In this section the estimates of §3 are applied to the study of exact controllability of the beam network in situations where the controls act through forces and bending moments applied at the joints of the system.

The controlled system is modeled by the system of equations

$$\begin{aligned}
 (4.1) \quad & \rho_i \ddot{u}_i - E_i A_i u_i'' = 0, \\
 & \rho_i \ddot{w}_i - K_i(\psi_i + w_i')' = 0, \\
 & I_{\rho_i} \ddot{\psi}_i - E_i I_i \psi_i'' + K_i(\psi_i + w_i') = 0, \quad 0 < x < \ell_i, \quad t > 0,
 \end{aligned}$$

$$(4.2) \quad u_i(N_k, t) = w_i(N_k, t) = \psi_i(N_k, t) = 0, \quad \forall i \in \mathcal{E}_k, \quad k \in I^D.$$

$$\begin{aligned}
 (4.3) \quad & E_i A_i u_i'(N_i, t) = f_{1i}(t), \\
 & K_i(\psi_i + w_i')(N_i, t) = f_{2i}(t), \\
 & E_i I_i \psi_i'(N_i, t) = f_{3i}(t), \quad i = 1, \dots, p,
 \end{aligned}$$

$$(4.4) \quad \mathbf{r}_i(N_k, t) = \mathbf{r}_j(N_k, t), \quad \psi_i(N_k, t) = \psi_j(N_k, t), \quad \forall i, j \in \mathcal{E}_k, \quad \forall k \in I_M^N,$$

and

$$(4.5) \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} [E_i A_i u_i' \mathbf{e}_i + K_i(\psi_i + w_i') \mathbf{e}_i^\perp + E_i I_i \psi_i' \mathbf{n}](N_k, t) = \mathbf{F}_k, \quad \forall k \in I_M.$$

The controls are

$$\begin{aligned}
 \mathbf{f}_i &= f_{1i} \mathbf{e}_i + f_{2i} \mathbf{e}_i^\perp + f_{3i} \mathbf{n}, \quad i = 1, \dots, p, \\
 \mathbf{F}_k &= (\mathbf{F}_k \cdot \mathbf{e}_i) \mathbf{e}_i + (\mathbf{F}_k \cdot \mathbf{e}_i^\perp) \mathbf{e}_i^\perp + (\mathbf{F}_k \cdot \mathbf{n}) \mathbf{n}, \quad k \in I_M^N
 \end{aligned}$$

(the latter are independent of i for $i \in \mathcal{E}_k$). The components of the controls in the $\mathbf{e}_i, \mathbf{e}_i^\perp$ plane represent forces acting at the corresponding joints, while the components along \mathbf{n} represent bending moments (torques) at the corresponding joints. Some of these controls may be inactive, i.e., equal to zero for $t > 0$. We denote by \mathbf{u} the vector of active controls. Then, for example, if the controls at the free multiple joints are inactive,

$$\mathbf{u}(t) = \bigotimes_{i=1}^p \mathbf{f}_i(t) \in U, \quad t > 0,$$

where

$$(4.6) \quad U = \prod_{i=1}^p U_i, \quad U_i = [\mathbf{e}_i] \bigoplus [\mathbf{e}_i^\perp] \bigoplus [\mathbf{n}].$$

Remark 4.1. In (4.3) and (4.5) the masses of the joints have not been taken into account. Accounting for the masses of the free multiple joints, in particular, would require adding $M_k \ddot{\mathbf{r}}_{i_k} + I_k \ddot{\psi}_{i_k} \mathbf{n}$ to the right-hand member in (4.5), where M_k denotes the mass of the k th joint and I_k its principal moment of inertia.

As is well known, the exact controllability problem for (4.1)–(4.5) may be formulated in terms of the *reachability problem*: for vanishing initial data

$$(4.7) \quad \mathbf{R}(0) = \dot{\mathbf{R}}(0) = 0,$$

determine the reachable states

$$\mathcal{R}_T = \{(\mathbf{R}(T), \dot{\mathbf{R}}(T)) | \mathbf{u} \in \mathcal{C}\},$$

where \mathcal{C} is the space of controls. In this work, \mathcal{C} will be taken to be

$$(4.8) \quad \mathcal{C} = L^2_{\text{loc}}(0, \infty; \mathbb{R}^q),$$

where q is the number of active (scalar) controls.

With \mathcal{C} given by (4.8), existence and uniqueness of solutions of (4.1)–(4.7) can easily be proved within a variational framework. To simplify the discussion a little, we assume that $I_S^D \neq \emptyset$, but this is inessential. Proceeding along the lines of §3, let $\hat{\mathbf{R}} \in V$ be a test function, form the $L^2(0, \ell_i; \mathbb{R}^3)$ scalar product of $\hat{\mathbf{R}}_i$ with (4.1) and sum over i . After some integrations by parts we obtain the variational equation

$$(4.9) \quad (\ddot{\mathbf{R}}, \hat{\mathbf{R}})_H + (\mathbf{R}, \hat{\mathbf{R}})_V = \sum_{i=1}^p \mathbf{f}_i \cdot \hat{\mathbf{R}}_i(N_i) + \sum_{k \in I_M^N} \mathbf{F}_k \cdot \hat{\mathbf{R}}_{i_k}(N_k)$$

where, as usual, i_k denotes any index in \mathcal{E}_k in accordance with the geometric node conditions.

Set

$$\mathbf{u} = \left(\bigotimes_{i=1}^p \mathbf{f}_i \right) \otimes \left(\bigotimes_{k \in I_M^N} \mathbf{F}_k \right) \in U,$$

where

$$U = \left(\prod_{i=1}^p U_i \right) \times \left(\prod_{k \in I_M^N} U_{i_k} \right)$$

and U_i is defined in (4.6). We have

$$\left| \sum_{i=1}^p \mathbf{f}_i \cdot \hat{\mathbf{R}}_i(N_i) + \sum_{k \in I_M^N} \mathbf{F}_k \cdot \hat{\mathbf{R}}_{i_k}(N_k) \right| \leq C \|\mathbf{u}\|_U \|\hat{\mathbf{R}}\|_V.$$

Therefore we may define $B \in \mathcal{L}(U, V')$ by

$$(4.10) \quad \langle B\mathbf{u}, \hat{\mathbf{R}} \rangle_V = \sum_{i=1}^p \mathbf{f}_i \cdot \hat{\mathbf{R}}_i(N_i) + \sum_{k \in I_M^N} \mathbf{F}_k \cdot \hat{\mathbf{R}}_{i_k}(N_k), \quad \forall \hat{\mathbf{R}} \in V,$$

where $\langle \cdot, \cdot \rangle_V$ denotes the V', V duality pairing. Let us note that, defining $B' \in \mathcal{L}(V, U)$ through $\langle B\mathbf{u}, \hat{\mathbf{R}} \rangle_V = \langle \mathbf{u}, B'\hat{\mathbf{R}} \rangle_U$, we have

$$B'\hat{\mathbf{R}} = \left(\bigotimes_{i=1}^p \hat{\mathbf{R}}_i(N_i) \right) \otimes \left(\bigotimes_{k \in I_M^N} \hat{\mathbf{R}}_{i_k}(N_k) \right).$$

From (4.9) and (4.10) we obtain the variational equation

$$(4.11) \quad \ddot{\mathbf{R}} + A\mathbf{R} = B\mathbf{u} \quad \text{in } V', \quad \mathbf{u} \in L^2_{\text{loc}}(0, \infty; U).$$

It follows from standard theory that (4.11), with initial data (4.7), has a unique solution with $(\mathbf{R}, \dot{\mathbf{R}}) \in C([0, T]; H \times V')$ for every $T > 0$. In fact, we may prove that $(\mathbf{R}, \dot{\mathbf{R}}) \in C([0, T]; V \times H)$. To do so, and to introduce some notation for later, we write (4.11) as a system in $H \times V'$ by introducing

$$\Phi = \begin{pmatrix} \mathbf{R} \\ \dot{\mathbf{R}} \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} 0 \\ B \end{pmatrix}.$$

We obtain

$$(4.12) \quad \dot{\Phi} = \mathcal{A}\Phi + \mathcal{B}\mathbf{u}, \quad \Phi(0) = 0.$$

The operator \mathcal{A} is skew-adjoint as an operator in $H \times V'$ with domain $V \times H$, and $\mathcal{B} \in \mathcal{L}(U, H \times V')$. The dual \mathcal{A}' of \mathcal{A} is the operator in $H \times V$ given by

$$\mathcal{A}' = \begin{pmatrix} 0 & A \\ -I & 0 \end{pmatrix}, \quad \text{Dom}(\mathcal{A}') = V \times D_A.$$

For each $u \in L^2(0, T; U) := \mathcal{U}$, the unique solution of (4.12) is given by

$$(\mathcal{S}_T \mathbf{u})(t) := \Phi(t) = \int_0^t e^{(t-s)\mathcal{A}} \mathcal{B}\mathbf{u}(s) ds, \quad 0 \leq t \leq T,$$

where $e^{t\mathcal{A}}, t \geq 0$, is the unitary group on $H \times V'$ generated by \mathcal{A} . We have that \mathcal{S}_T is a bounded linear map from \mathcal{U} into $C([0, T]; H \times V')$. In fact, \mathcal{S}_T is bounded from \mathcal{U} into $C([0, T]; V \times H)$. To prove this, let $\hat{\Phi}^0 \in H \times V$ and form

$$(4.13) \quad \left\langle \int_0^\tau e^{(\tau-s)\mathcal{A}} \mathcal{B}\mathbf{u}(s) ds, \hat{\Phi}^0 \right\rangle_{H \times V} = \int_0^\tau (\mathbf{u}(s), \mathcal{B}' e^{(\tau-s)\mathcal{A}'} \hat{\Phi}^0)_U ds, \quad 0 \leq \tau \leq T,$$

where

$$\mathcal{B}' = \begin{pmatrix} 0 \\ B' \end{pmatrix} \in \mathcal{L}(H \times V, U).$$

Set $\hat{\Phi}(t) = e^{(\tau-t)\mathcal{A}'} \hat{\Phi}^0$. Then

$$(4.14) \quad \dot{\hat{\Phi}}(t) = -\mathcal{A}'\hat{\Phi}(t), \quad 0 \leq t < \tau, \quad \hat{\Phi}(\tau) = \hat{\Phi}^0.$$

Writing $\hat{\Phi} = (\hat{\mathbf{R}}_1, \hat{\mathbf{R}})$, $\hat{\Phi}^0 = (\hat{\mathbf{R}}^1, \hat{\mathbf{R}}^0)$, (4.14) signifies that

$$\ddot{\hat{\mathbf{R}}} + A\hat{\mathbf{R}} = 0, \quad \hat{\mathbf{R}}_1 = \dot{\hat{\mathbf{R}}},$$

$$\hat{\mathbf{R}}(\tau) = \hat{\mathbf{R}}^0, \quad \dot{\hat{\mathbf{R}}}(\tau) = \hat{\mathbf{R}}^1.$$

From the inequality (3.16), applied to $\hat{\mathbf{R}}$, we have

$$\int_0^\tau |\mathcal{B}'\hat{\Phi}(t)|_U^2 dt = \int_0^\tau |B'\hat{\mathbf{R}}(t)|_U^2 dt \leq C(\tau+1)(\|\hat{\mathbf{R}}^0\|_H^2 + \|\hat{\mathbf{R}}^1\|_{V'}^2) = C(\tau+1)\|\hat{\Phi}^0\|_{V' \times H}^2.$$

Use of this inequality in the right-hand side of (4.13) allows us to conclude that

$$\sup_{0 \leq \tau \leq T} \left| \left\langle \int_0^\tau e^{(\tau-s)\mathcal{A}} \mathbf{B} \mathbf{u}(s) ds, \hat{\Phi}^0 \right\rangle_{H \times V} \right| \leq C \sqrt{T+1} \|\mathbf{u}\|_{\mathcal{U}} \|\hat{\Phi}^0\|_{V' \times H}, \quad \forall \hat{\Phi}^0 \in H \times V.$$

It follows that

$$\Phi = S_T \mathbf{u} = (\mathbf{R}, \dot{\mathbf{R}}) \in L^\infty(0, T; V \times H).$$

We may then pass from L^∞ to C by a standard argument. We have proved the following proposition.

PROPOSITION 4.1. *For an arbitrary control $\mathbf{u} \in L^2_{loc}(0, \infty; \mathbb{R}^{3m})$, the system (4.1)–(4.7) has a unique finite energy solution.*

We now know that $(\mathbf{R}(T), \dot{\mathbf{R}}(T)) \in V \times H$ for every $T > 0$, whenever $\mathbf{u} \in \mathcal{C}$. We want to determine under what conditions it is true that *every element* in $V \times H$ is also in \mathcal{R}_T , i.e., S_T maps \mathcal{U} onto $H \times V$, where $S_T \mathbf{u} := (S_T \mathbf{u})(T)$. The validity of the statement $\mathcal{R}_T = V \times H$ depends on the number and placements of the active controls and these, in turn, depend on the particular beam configuration.

4.1. Controllability from the simple nodes. We begin by considering the case where the active controls are placed only at simple free nodes. Here we have U defined by (4.6) and the control operator $B \in \mathcal{L}(U, V')$ and its dual B' in this situation are defined, respectively, by

$$\langle B \mathbf{u}, \hat{\mathbf{R}} \rangle_V = \sum_{i=1}^p \mathbf{f}_i \cdot \hat{\mathbf{R}}_i(N_i), \quad \forall \hat{\mathbf{R}} \in V,$$

$$B' \hat{\mathbf{R}} = \bigotimes_{i=1}^p \hat{\mathbf{R}}_i(N_i), \quad \forall \hat{\mathbf{R}} \in V.$$

Let $S'_T \in \mathcal{L}(V' \times H, \mathcal{U})$ denote the dual of S_T ; it is defined by

$$\langle \hat{\Phi}^0, S_T \mathbf{u} \rangle_{V \times H} = (S'_T \hat{\Phi}^0, \mathbf{u})_{\mathcal{U}}, \quad \forall \hat{\Phi}^0 \in V' \times H, \forall \mathbf{u} \in \mathcal{U}.$$

Since

$$\langle \hat{\Phi}^0, S_T \mathbf{u} \rangle_{V \times H} = \int_0^T (\mathbf{u}(s), B' e^{(T-s)\mathcal{A}'} \hat{\Phi}^0)_U ds,$$

we have

$$(S'_T \hat{\Phi}^0)(s) = B' e^{(T-s)\mathcal{A}'} \hat{\Phi}^0.$$

The assertion $\text{Range}(S_T) = V \times H$ is equivalent to

$$\|S'_T \hat{\Phi}^0\|_{\mathcal{U}} \geq c_T \|\hat{\Phi}^0\|_{V' \times H}, \quad \forall \hat{\Phi}^0 \in V' \times H,$$

for some constant $c_T > 0$, that is,

$$(4.15) \quad \int_0^T |B' \hat{\mathbf{R}}(t)|_{\mathcal{U}}^2 dt \geq c_T (\|\hat{\mathbf{R}}^0\|_H^2 + \|\hat{\mathbf{R}}^1\|_{V'}^2), \quad \forall (\hat{\mathbf{R}}^0, \hat{\mathbf{R}}^1) \in H \times V',$$

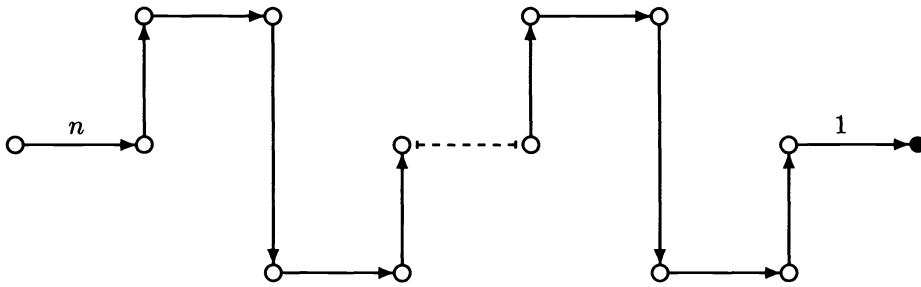


FIG. 3. An example of an exactly controllable network. Beam n is clamped at its simple node and controls are applied at the free node of beam 1.

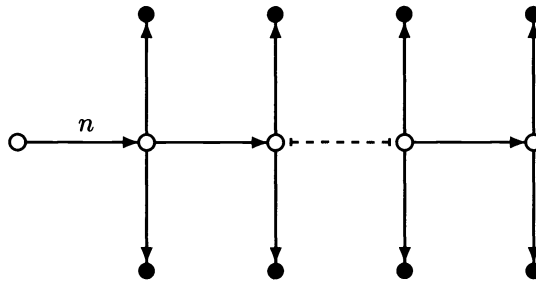


FIG. 4. An example of an exactly controllable network. Beam n is clamped at its simple node and controls are applied at all other simple nodes.

where $\hat{\mathbf{R}}$ is the solution of

$$(4.16) \quad \ddot{\hat{\mathbf{R}}} + A\hat{\mathbf{R}} = 0, \quad 0 \leq t < T, \quad \hat{\mathbf{R}}(T) = \hat{\mathbf{R}}^0, \quad \dot{\hat{\mathbf{R}}}(T) = \hat{\mathbf{R}}^1.$$

Sufficient conditions for the validity of (4.15) were given in §3. When (4.15) holds, the control \mathbf{u}_0 of minimum norm in \mathcal{U} that drives the initial data (4.7) to a prescribed state $(\mathbf{R}^0, \mathbf{R}^1) \in V \times H$ at time T is given by $\mathbf{u}_0 = B'\hat{\mathbf{R}}$, where $\hat{\mathbf{R}}$ is the solution of (4.16) with final data $(\hat{\mathbf{R}}^1, \hat{\mathbf{R}}^0) = (S_T S_T')^{-1}(\mathbf{R}^0, \mathbf{R}^1)$.

Example 4.1. Consider the following beam network shown in Fig. 3. In this figure, beam n is clamped at its simple node, a control \mathbf{f}_1 is applied at the simple node of beam 1 and the multiple nodes of the system are free. The angle between successive beams (determined by $\mathbf{e}_i \cdot \mathbf{e}_{i+1}$) may be chosen arbitrarily in the interval $[0, \pi]$. Based on the discussion of §3, the network exhibited in Fig. 3 is exactly controllable, i.e., $\mathcal{R}_T = V \times H$ if T is sufficiently large. This result may be compared to a result of Chen et al. [1], where uniform stabilization in finite energy space of a network of *serially* connected (i.e., $\mathbf{e}_i = \mathbf{e}_1$ for all i) Euler–Bernoulli beams was established (equivalent to $\mathcal{R}_T = V \times H$ for T sufficiently large), but only under a monotonicity requirement on the stiffnesses of successive beams.

Example 4.2. Another exactly controllable beam network is the H-shaped configuration shown in Fig. 4. Beam n is clamped at its simple node, controls (indicated by solid circles) are applied at all other simple nodes and the multiple nodes are free.

The angle of $\pi/2$ between horizontal and vertical beams is drawn only for convenience; any angle in $(0, \pi)$ is admissible.

Example 4.3. Figure 2 is an example of a network that we cannot prove is exactly controllable when controls act only at the simple nodes. In fact, the discussion of the next subsection shows that the network in Fig. 2 may not be even approximately controllable in all cases, even when controls are active at each node of the network.

4.2. Controllability with controls at the multiple nodes. In this subsection we wish to consider the question of controllability in situations where the network may not be exactly controllable by means of controls acting only at the free simple nodes. Such will be the case, for example, when $I_S^N = \emptyset$; or if the network contains a closed loop, as in Fig. 2. Although we are unable at the present time to give a definitive controllability analysis in these situations, we will indicate, through examples, some of the anomalies that can occur. What these examples show is that there are certain beam configurations (such as in Fig. 2) that may support nontrivial time periodic wave motions that cannot be controlled through controls placed at the available nodes of the system. On the other hand, such situations are certainly “nongeneric”; they occur only when the elastic parameters and beam lengths are related in very specific ways.

Only *approximate controllability* will be considered in what follows. The system (4.11) is approximately controllable at time T if $\bar{\mathcal{R}}_T = V \times H$ or, equivalently, if the range of S_T is dense in $V \times H$. The latter is the same as the assertion that the kernel of S'_T contains only the zero element of $V' \times H$. From the definition of S'_T , this is the same as saying that the only finite energy solution of the problem

$$(4.17) \quad \ddot{\mathbf{R}}(t) + A\mathbf{R}(t) = 0, \quad B'\mathbf{R}(t) = 0, \quad 0 < t < T,$$

is $\mathbf{R}(t) \equiv 0$.

From the definition of the control operator B , the condition $B'\mathbf{R} = 0$ means that

$$\mathbf{R}_i = u_i \mathbf{e}_i + w_i \mathbf{e}_i^\perp + \psi_i \mathbf{n} = 0,$$

at each simple, free node N_i and at each free multiple node N_k (for all $i \in \mathcal{E}_k$) where B has a nonzero component. Our main objective in what follows is to demonstrate that for certain beam networks, (4.17) admits nontrivial periodic solutions.

In order that the computations not obscure the main ideas, we will first consider networks of vibrating *strings*. Modeling and controllability of such networks have been studied by Schmidt [9] and we adopt the following model directly from [9]. With the same notation as above, the controlled model is

$$(4.18) \quad \ddot{\mathbf{r}}_i - K_i \mathbf{r}_i'' = 0, \quad i = 1, \dots, n,$$

where K_i is a 2×2 diagonal matrix with respect to the $\mathbf{e}_i, \mathbf{e}_i^\perp$ basis having real entries p_i^2, q_i^2 on the diagonal;

$$(4.19) \quad \mathbf{r}_i(N_i, t) = 0, \quad i \in I_S^D;$$

$$(4.20) \quad K_i \mathbf{r}_i'(N_i, t) = \mathbf{f}_i, \quad i = 1, \dots, p;$$

$$(4.21) \quad \mathbf{r}_i(N_k, t) = \mathbf{r}_j(N_k, t), \quad \forall i, j \in \mathcal{E}_k, \quad k \in I_M,$$

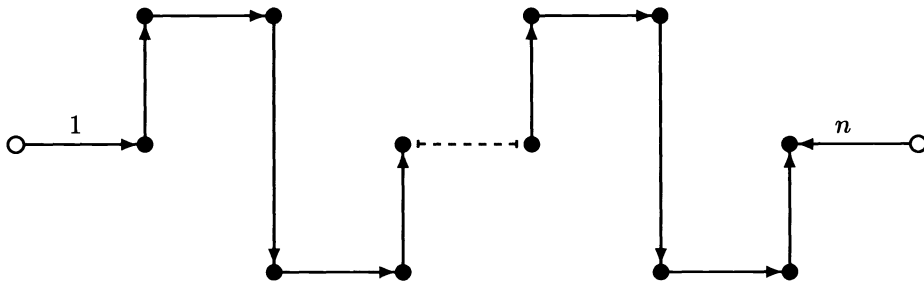


FIG. 5. Beams 1 and n are clamped at their simple nodes and controls are applied at all multiple nodes. Such a network is not approximately controllable, in general.

which are the geometric node conditions; and

$$(4.22) \quad \sum_{i \in \mathcal{E}_k} \varepsilon_{ik} (p_i^2 u'_i \mathbf{e}_i + q_i^2 w'_i \mathbf{e}_i^\perp)(N_k, t) = \mathbf{F}_k, \quad k \in I_M,$$

which are the dynamic node conditions. The dependent variables u_i, w_i represent longitudinal and transverse displacements respectively, along the i th string.

For the above string network, (4.17) consists of (4.18)–(4.22) with $\mathbf{f}_i = 0$, $\mathbf{F}_k = 0$, together with

$$(4.23) \quad \mathbf{r}_i = 0$$

at each simple free node N_i and at each multiple node N_k , for all $i \in \mathcal{E}_k$, where a control is active, i.e., where the control operator has a nonzero component. We want to look for eigenvalues of this system, so we set

$$\mathbf{r}_i(x, t) = e^{\sqrt{-1}\lambda t} \mathbf{r}_i(x), \quad i = 1, \dots, n,$$

and consider

$$(4.24) \quad K_i \mathbf{r}_i'' + \lambda^2 \mathbf{r}_i = 0, \quad i = 1, \dots, n,$$

$$\mathbf{r}_i(N_i) = 0, \quad i \in I_S^D, \quad K_i \mathbf{r}_i'(N_i) = 0, \quad i = 1, \dots, p,$$

$$\mathbf{r}_i(N_k) = \mathbf{r}_j(N_k), \quad i, j \in \mathcal{E}_k, \quad k \in I_M,$$

$$\sum_{i \in \mathcal{E}_k} \varepsilon_{ik} (p_i^2 u'_i \mathbf{e}_i + q_i^2 w'_i \mathbf{e}_i^\perp)(N_k) = 0, \quad k \in I_M.$$

In addition, (4.23) must hold at appropriate nodes, as described above.

The following two examples illustrate situations where the above eigenvalue problem has nontrivial solutions.

Example 4.4. Consider the network of Example 4.1, but with both simple nodes clamped and all multiple nodes controlled, as illustrated in Fig. 5. The eigenvalue

problem in this case consists of (4.24) together with

$$(4.25) \quad \mathbf{r}_i(0) = \mathbf{r}_i(\ell_i) = 0, \quad i = 1, \dots, n,$$

$$(4.26) \quad p_i^2 u'_i(\ell_i) \mathbf{e}_i + q_i^2 w'_i(\ell_i) \mathbf{e}_i^\perp = p_{i+1}^2 u'_{i+1}(0) \mathbf{e}_{i+1} + q_{i+1}^2 w'_{i+1}(0) \mathbf{e}_{i+1}^\perp, \\ i = 1, \dots, n-2,$$

$$(4.27) \quad p_{n-1}^2 u'_{n-1}(\ell_{n-1}) \mathbf{e}_{n-1} + q_{n-1}^2 w'_{n-1}(\ell_{n-1}) \mathbf{e}_{n-1}^\perp \\ = -[p_n^2 u'_n(\ell_n) \mathbf{e}_n + q_n^2 w'_n(\ell_n) \mathbf{e}_n^\perp].$$

From (4.24), (4.25) we must have

$$u_i(x) = u_i^0 \sin(\lambda/p_i)x, \quad w_i(x) = w_i^0 \sin(\lambda/q_i)x,$$

where

$$(4.28) \quad \sin(\lambda/p_i)\ell_i = \sin(\lambda/q_i)\ell_i = 0, \quad i = 1, \dots, n.$$

Equations (4.28) admit an infinite sequence of solutions $\lambda^{(j)} \rightarrow \infty$ provided the numbers $\ell_i/p_i, \ell_i/q_i$, $i = 1, \dots, n$, are comeasurable, i.e., their ratios are rational. In particular, p_i and q_i must then be comeasurable for each i .

Let λ be a nonzero solution of (4.28) such that

$$\cos(\lambda/p_i)\ell_i = \cos(\lambda/q_i)\ell_i = 1, \quad i = i, \dots, n.$$

Conditions (4.26), (4.27) require that

$$p_i u_i^0 \mathbf{e}_i + q_i w_i^0 \mathbf{e}_i^\perp = p_{i+1} u_{i+1}^0 \mathbf{e}_{i+1} + q_{i+1} w_{i+1}^0 \mathbf{e}_{i+1}^\perp, \quad i = 1, \dots, n-2, \\ p_{n-1} u_{n-1}^0 \mathbf{e}_{n-1} + q_{n-1} w_{n-1}^0 \mathbf{e}_{n-1}^\perp = -(p_n u_n^0 \mathbf{e}_n + q_n w_n^0 \mathbf{e}_n^\perp).$$

With $\mathbf{r}_1^0 = u_1^0 \mathbf{e}_1 + w_1^0 \mathbf{e}_1^\perp$ arbitrarily given, this last system uniquely determines \mathbf{r}_i^0 for $i = 2, \dots, n$. Therefore the network of Fig. 5 is not approximately controllable when the above conditions on the relative lengths and wave speeds of the various strings are met. On the other hand, in general (4.28) will admit only the trivial solution $\lambda = 0$. One may prove (although we shall not do so here) that the network is then approximately controllable, but we do not know if it is exactly controllable in such cases.

A similar phenomenon may occur in a network of strings containing a closed loop.

Example 4.5. Consider a closed loop consisting of n strings and n nodes, with controls at each node (see Fig. 6). Existence of periodic solutions $e^{\sqrt{-1}\lambda t} \mathbf{r}_i(x)$ of (4.17) for such a network requires that (4.25) hold, that

$$(4.29) \quad p_i^2 u'_i(\ell_i) \mathbf{e}_i + q_i^2 w'_i(\ell_i) \mathbf{e}_i^\perp = p_{i+1}^2 u'_{i+1}(0) \mathbf{e}_{i+1} + q_{i+1}^2 w'_{i+1}(0) \mathbf{e}_{i+1}^\perp, \\ i = 1, \dots, n-1,$$

and that

$$(4.30) \quad p_n^2 u'_n(\ell_n) \mathbf{e}_n + q_n^2 w'_n(\ell_n) \mathbf{e}_n^\perp = p_1^2 u'_1(0) \mathbf{e}_1 + q_1^2 w'_1(0) \mathbf{e}_1^\perp.$$

Assume that p_i, q_i, ℓ_i are such that (4.28) admits a nontrivial solution λ with

$$\cos(\lambda/p_i)\ell_i = \cos(\lambda/q_i)\ell_i = 1.$$

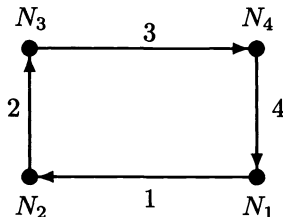


FIG. 6. Networks containing closed loops are not approximately controllable from their nodes, in general, regardless of the number of controlled nodes.

Given \mathbf{r}_1^0 , equations (4.29) then uniquely determine \mathbf{r}_i^0 , $i = 2, \dots, n$, in terms of \mathbf{r}_1^0 . The additional “compatibility condition” (4.30) requires that

$$p_1 u_1^0 \mathbf{e}_1 + q_1 w_1^0 \mathbf{e}_1^\perp = p_n u_n^0 \mathbf{e}_n + q_n w_n^0 \mathbf{e}_n^\perp,$$

which is easily seen to be a consequence of (4.29). Thus, under the above special conditions on string lengths and wave speeds, Fig. 6 is not an approximately controllable configuration. The same will be true even if Fig. 6 has other strings attached to it, as in Fig. 2. If all nodes in Fig. 2 are controlled, there can still be nontrivial periodic solutions; these will consist of a periodic wave traveling around the closed loop, as above, with the solution being identically zero in the appendages.

Now let us consider the configurations of Examples 4.4 and 4.5 in the context of Timoshenko beams. As is seen from the string examples, the main issue is whether the different beams of the network, subject to clamped boundary conditions, can have a common eigenvalue and, if so, whether corresponding eigenfunctions can be chosen to satisfy the given node conditions of the problem.

Consider first the configuration of Fig. 5. We are looking for periodic solutions of (4.17) of the form $\mathbf{r}_i(x, t) = e^{\sqrt{-1}\lambda t} \mathbf{r}_i(x)$, $\psi_i(x, t) = e^{\sqrt{-1}\mu t} \psi_i(x)$, which reflects the possibility of different wave speeds for elastic and shear waves. The triple u_i, w_i, ψ_i , $0 \leq x \leq \ell_i$, must then satisfy

$$\begin{aligned} (4.31) \quad & u_i'' + s_i^2 \lambda^2 u_i = 0, \quad s_i^2 = \rho_i / E_i A_i, \\ & (\psi_i + w_i')' + p_i^2 \lambda^2 w_i = 0, \quad p_i^2 = \rho_i / K_i, \\ & \psi_i'' + q_i^2 \mu^2 \psi_i - r_i^2 (\psi_i + w_i') = 0, \quad q_i^2 = I_{\rho_i} / E_i I_i, \quad r_i^2 = K_i / E_i I_i; \end{aligned}$$

$$(4.32) \quad u_i(0) = u_i(\ell_i) = w_i(0) = w_i(\ell_i) = \psi_i(0) = \psi_i(\ell_i) = 0, \quad i = 1, \dots, n;$$

$$(4.33) \quad \begin{aligned} E_i I_i \psi_i'(\ell_i) &= E_{i+1} I_{i+1} \psi_{i+1}'(0), \quad i = 1, \dots, n-2, \\ E_{n-1} I_{n-1} \psi_{n-1}'(\ell_{n-1}) &= E_n I_n \psi_n'(\ell_n); \end{aligned}$$

$$(4.34) \quad \begin{aligned} E_i A_i u_i'(\ell_i) \mathbf{e}_i + K_i w_i'(\ell_i) \mathbf{e}_i^\perp &= E_{i+1} A_{i+1} u_{i+1}'(0) \mathbf{e}_{i+1} + K_{i+1} w_{i+1}'(0) \mathbf{e}_{i+1}^\perp, \\ &\quad i = 1, \dots, n-2, \\ E_{n-1} A_{n-1} u_{n-1}'(\ell_{n-1}) \mathbf{e}_{n-1} + K_{n-1} w_{n-1}'(\ell_{n-1}) \mathbf{e}_{n-1}^\perp \\ &= E_n A_n u_n'(\ell_n) \mathbf{e}_n + K_n w_n'(\ell_n) \mathbf{e}_n^\perp \end{aligned}$$

From the first equation in (4.31) and (4.32) we must have

$$(4.35) \quad u_i(x) = u_i^0 \sin \lambda s_i x$$

with

$$(4.36) \quad \sin \lambda s_i \ell_i = 0, \quad i = 1, \dots, n.$$

The equations in w_i and ψ_i in (4.31) may be uncoupled as follows. We have

$$0 = w_i''' + \psi_i'' + p_i^2 \lambda^2 w_i' = w_i''' + r_i^2 (\psi_i + w_i') - q_i^2 \mu^2 \psi_i + p_i^2 \lambda^2 w_i'$$

which yields, upon differentiation,

$$\begin{aligned} 0 &= w_i'''' + r_i^2 (\psi_i + w_i')' - q_i^2 \mu^2 \psi_i' + p_i^2 \lambda^2 w_i'' \\ &= w_i'''' - r_i^2 p_i^2 \lambda^2 w_i - q_i^2 \mu^2 (-p_i^2 \lambda^2 w_i - w_i'') + p_i^2 \lambda^2 w_i'', \end{aligned}$$

that is,

$$(4.37) \quad w_i'''' + (p_i^2 \lambda^2 + q_i^2 \mu^2) w_i'' + p_i^2 \lambda^2 (q_i^2 \mu^2 - r_i^2) w_i = 0.$$

We define

$$\sigma_i^\pm(\lambda, \mu) = \frac{1}{2} \left\{ p_i^2 \lambda^2 + q_i^2 \mu^2 \pm [(p_i^2 \lambda^2 - q_i^2 \mu^2)^2 + 4p_i^2 r_i^2 \lambda^2]^{1/2} \right\},$$

and note that $\sigma_i^\pm(\lambda, \mu) > 0$ if $q_i^2 \mu^2 > r_i^2$, which we shall assume. The roots of the characteristic equation of (4.37) are $\pm \sqrt{-1} [\sigma_i^\pm(\lambda, \mu)]^{1/2}$, so that the general solution of (4.37) is

$$(4.38) \quad w_i(x) = a_i \sin x \sqrt{\sigma_i^+} + b_i \cos x \sqrt{\sigma_i^+} + c_i \sin x \sqrt{\sigma_i^-} + d_i \cos x \sqrt{\sigma_i^-}.$$

Upon setting $\psi_i' = -w_i'' - p_i^2 \lambda^2 w_i$ and integrating, we find that

$$(4.39) \quad \begin{aligned} \psi_i(x) &= \psi_i^0 + \frac{\sigma_i^+ - p_i^2 \lambda^2}{\sqrt{\sigma_i^+}} \left(b_i \sin x \sqrt{\sigma_i^+} - a_i \cos x \sqrt{\sigma_i^+} \right) \\ &\quad + \frac{\sigma_i^- - p_i^2 \lambda^2}{\sqrt{\sigma_i^-}} \left(d_i \sin x \sqrt{\sigma_i^-} - c_i \cos x \sqrt{\sigma_i^-} \right). \end{aligned}$$

Since then

$$\psi_i'' + (q_i^2 \mu^2 - r_i^2) \psi_i - r_i^2 w_i' = (q_i^2 \mu^2 - r_i^2) \psi_i^0,$$

to satisfy the third equation in (4.31) we must set $\psi_i^0 = 0$.

The boundary conditions $w_i(0) = \psi_i(0) = 0$ require that

$$b_i = -d_i, \quad c_i = -a_i \frac{\sqrt{\sigma_i^-} (\sigma_i^+ - p_i^2 \lambda^2)}{\sqrt{\sigma_i^+} (\sigma_i^- - p_i^2 \lambda^2)},$$

and therefore

$$w_i(x) = a_i \left(\sin x \sqrt{\sigma_i^+} - \frac{\sqrt{\sigma_i^-} (\sigma_i^+ - p_i^2 \lambda^2)}{\sqrt{\sigma_i^+} (\sigma_i^- - p_i^2 \lambda^2)} \sin x \sqrt{\sigma_i^-} \right) + b_i \left(\cos x \sqrt{\sigma_i^+} - \cos x \sqrt{\sigma_i^-} \right),$$

$$\begin{aligned} \psi_i(x) = & b_i \frac{\sigma_i^+ - p_i^2 \lambda^2}{\sqrt{\sigma_i^+}} \left(\sin x \sqrt{\sigma_i^+} - \frac{\sqrt{\sigma_i^+}(\sigma_i^- - p_i^2 \lambda^2)}{\sqrt{\sigma_i^-}(\sigma_i^+ - p_i^2 \lambda^2)} \sin x \sqrt{\sigma_i^-} \right) \\ & - a_i \frac{\sigma_i^+ - p_i^2 \lambda^2}{\sqrt{\sigma_i^+}} \left(\cos x \sqrt{\sigma_i^+} - \cos x \sqrt{\sigma_i^-} \right). \end{aligned}$$

The conditions $w_i(\ell_i) = \psi_i(\ell_i) = 0$ will yield nontrivial solution a_i, b_i if and only if the matrix

$$(4.40) \quad \begin{pmatrix} \sin \ell_i \sqrt{\sigma_i^+} - \frac{\sqrt{\sigma_i^-}(\sigma_i^+ - p_i^2 \lambda^2)}{\sqrt{\sigma_i^+}(\sigma_i^- - p_i^2 \lambda^2)} \sin \ell_i \sqrt{\sigma_i^-} & \cos \ell_i \sqrt{\sigma_i^+} - \cos \ell_i \sqrt{\sigma_i^-} \\ \cos \ell_i \sqrt{\sigma_i^-} - \cos \ell_i \sqrt{\sigma_i^+} & \sin \ell_i \sqrt{\sigma_i^+} - \frac{\sqrt{\sigma_i^+}(\sigma_i^- - p_i^2 \lambda^2)}{\sqrt{\sigma_i^-}(\sigma_i^+ - p_i^2 \lambda^2)} \sin \ell_i \sqrt{\sigma_i^-} \end{pmatrix}$$

is rank deficient, that is, if and only if

$$(4.41) \quad \begin{aligned} & \frac{\sigma_i^- (\sigma_i^+ - p_i^2 \lambda^2)^2 + \sigma_i^+ (\sigma_i^- - p_i^2 \lambda^2)^2}{\sqrt{\sigma_i^+ \sigma_i^-} (\sigma_i^+ - p_i^2 \lambda^2) (\sigma_i^- - p_i^2 \lambda^2)} \sin \ell_i \sqrt{\sigma_i^+} \sin \ell_i \sqrt{\sigma_i^-} \\ & - 2 \cos \ell_i \sqrt{\sigma_i^+} \cos \ell_i \sqrt{\sigma_i^-} + 2 = 0, \quad i = 1, \dots, n. \end{aligned}$$

Therefore, the spectrum of (4.31), (4.32) is determined by (4.36) and (4.41).

In analogy with string networks considered above, we expect that if (4.36), (4.41) admit a nontrivial solution $\bar{\lambda}, \bar{\mu}$, then nontrivial values of a_i, b_i, u_i^0 , $i = 1, \dots, n$, can be chosen so that the dynamic node conditions (4.33), (4.34) are satisfied. However, this will not be the case, in general. For suppose the above matrix has rank one. Then the parameter a_i will be uniquely determined in terms of b_i , $i = 1, \dots, n$. The parameters u_i^0, b_1 must be completely free (since the problem is homogeneous). Once u_i^0, b_i are known, (4.33) and (4.34) yield *three* linear equations for the *two* parameters u_{i+1}^0, b_{i+1} . This overdetermined system will not, in general, have a solution, unless all u_i^0, b_i are zero.

Therefore, to obtain a nontrivial solution of (4.31)–(4.34), we assume that the matrix above has rank zero, i.e.,

$$(4.42) \quad \cos \ell_i \sqrt{\sigma_i^+} = \cos \ell_i \sqrt{\sigma_i^-},$$

$$(4.43) \quad \begin{aligned} \sin \ell_i \sqrt{\sigma_i^+} &= \frac{\sqrt{\sigma_i^-}(\sigma_i^+ - p_i^2 \lambda^2)}{\sqrt{\sigma_i^+}(\sigma_i^- - p_i^2 \lambda^2)} \sin \ell_i \sqrt{\sigma_i^-} \\ &= \frac{\sqrt{\sigma_i^+}(\sigma_i^- - p_i^2 \lambda^2)}{\sqrt{\sigma_i^-}(\sigma_i^+ - p_i^2 \lambda^2)} \sin \ell_i \sqrt{\sigma_i^-}. \end{aligned}$$

Both (4.42) and (4.43) will be satisfied for some λ, μ if

$$(4.44) \quad \ell_i^2 \sigma_i^+ (\lambda, \mu) = 4M_i^2 \pi^2, \quad \ell_i^2 \sigma_i^- (\lambda, \mu) = 4m_i^2 \pi^2$$

for integers M_i, m_i with $M_i > m_i$. From (4.44) we have

$$(4.45) \quad \begin{aligned} p_i^2 \lambda^2 + q_i^2 \mu^2 &= \sigma_i^+(\lambda, \mu) + \sigma_i^-(\lambda, \mu) = \frac{4(M_i^2 + m_i^2)\pi^2}{\ell_i^2}, \\ p_i^2 \lambda^2 (q_i^2 \mu^2 - r_i^2) &= \sigma_i^+(\lambda, \mu) \sigma_i^-(\lambda, \mu) = \frac{16m_i^2 M_i^2 \pi^4}{\ell_i^4}, \quad i = 1, \dots, n. \end{aligned}$$

Thus a spectral point λ, μ must satisfy (4.36) and (4.45).

If the numbers $\ell_i s_i$ are comeasurable, (4.36) has a sequence of solutions $\lambda^{(j)} \rightarrow \infty$. We select these so that $\cos \lambda^{(j)} s_i \ell_i = 1$ for $i = 1, \dots, n$. Then $\lambda^{(j)}$ may be expressed as

$$(4.46) \quad \lambda^{(j)} = 2j\pi \frac{n_i}{\ell_i s_i}$$

for some integer n_i . (We have $n_i = \ell_i s_i$ if $\ell_i s_i$ is an integer.) Substitution of $\lambda^{(j)}$ into (4.45) leads to $2n$ equations for μ , namely,

$$(4.47) \quad \mu^2 = \frac{4\pi^2}{\ell_i^2 q_i^2} \left[(M_i^2 + m_i^2) - \frac{j^2 n_i^2 p_i^2}{s_i^2} \right],$$

$$(4.48) \quad \mu^2 = \frac{1}{q_i^2} \left[\frac{4\pi^2 m_i^2 M_i^2 s_i^2}{j^2 \ell_i^2 n_i^2 p_i^2} + r_i^2 \right], \quad i = 1, \dots, n.$$

Therefore, existence of a spectral point requires the right-hand sides of (4.47) and (4.48) to be equal and independent of $i \in [1, \dots, n]$ for some choices of the integers m_i, M_i, j . These requirements restrict the quantities $p_i^2, q_i^2, r_i^2, s_i^2$. They correspond to the restriction in the string network case that the wave speeds p_i, q_i be comeasurable for each i . In particular, if all the beams are identical, the condition for the existence of a spectral point is (dropping the subscripts i)

$$(4.49) \quad \frac{4\pi^2}{\ell^2} \left(M^2 + m^2 - \frac{j^2 n^2 p^2}{s^2} \right) = \frac{4\pi^2 m^2 M^2 s^2}{j^2 \ell^2 n^2 p^2} + r^2$$

for some positive integers j, m, M . In this case there are no restrictions on q^2 . If also ℓs is an integer (so that $n = \ell s$), then (4.49) is a restriction only between p^2 and r^2 .

When the above conditions for the existence of a spectral point are met, the parameters a_i, b_i, u_i^0 are completely free. Equations (4.33) and (4.34) then uniquely determine $a_{i+1}, b_{i+1}, u_{i+1}^0$ in terms of a_i, b_i, u_i^0 . In fact, we find that (4.33) is

$$(4.50) \quad \frac{E_i I_i (M_i^2 - m_i^2)}{\ell_i^2} b_i = \frac{E_{i+1} I_{i+1} k_{i+1} (M_{i+1}^2 - m_{i+1}^2)}{\ell_{i+1}^2} b_{i+1}, \quad i = 1, \dots, n-1,$$

and (4.34) is

$$(4.51) \quad \begin{aligned} s_i \bar{\lambda} E_i A_i u_i^0 \mathbf{e}_i + \frac{2K_i (m_i + M_i) \pi}{\ell_i} a_i \mathbf{e}_i^\perp &= s_{i+1} \bar{\lambda} E_{i+1} A_{i+1} u_{i+1}^0 \mathbf{e}_{i+1} \\ &+ \frac{2K_{i+1} (m_{i+1} + M_{i+1}) \pi}{\ell_{i+1}} a_{i+1} \mathbf{e}_{i+1}^\perp, \quad i = 1, \dots, n-1. \end{aligned}$$

The a_i, b_i, u_i^0 will all be nonzero if a_1, b_1, u_1^0 are nonzero. It follows that, in the special circumstances described above, the Timoshenko beam network of Fig. 5 is not approximately controllable. However, this is obviously an anomalous situation.

A similar analysis may be applied to a closed-loop beam network such as that illustrated in Fig. 6. The only difference is that the node conditions (4.33)–(i) and (4.34)–(i) must hold for $i = 1, \dots, n-1$, and the “compatibility conditions”

$$\begin{aligned} E_n I_n \psi'_n(\ell_n) &= E_1 I_1 \psi'_1(0), \\ E_n A_n u'_n(\ell_n) \mathbf{e}_n + K_n w'_n(\ell_n) \mathbf{e}_n^\perp &= E_1 A_1 u'_1(0) \mathbf{e}_1 + K_1 w'_1(0) \mathbf{e}_1^\perp, \end{aligned}$$

must be satisfied. The former uniquely determine a_i, b_i, u_i^0 , $i = 2, \dots, n$, in terms of a_1, b_1, u_1^0 when the conditions described above are met; these values are given by (4.50) and (4.51). The compatibility conditions are

$$\begin{aligned} \frac{E_n I_n (M_n^2 - m_n^2)}{\ell_n^2} b_n &= \frac{E_1 I_1 (M_1^2 - m_1^2)}{\ell_1^2} b_1, \\ s_n \bar{\lambda} E_n A_n u_n^0 \mathbf{e}_n + \frac{2K_n (m_n + M_n) \pi}{\ell_n} a_n \mathbf{e}_n^\perp &= s_1 \bar{\lambda} E_1 A_1 u_1^0 \mathbf{e}_1 + \frac{2K_1 (m_1 + M_1) \pi}{\ell_1} a_1 \mathbf{e}_1^\perp, \end{aligned}$$

which are easily seen to be consequences of (4.50) and (4.51).

5. Asymptotic stability of beam networks. In this section we consider the asymptotic stability of the control system (4.10) under (possibly nonlinear) feedback controls of the form

$$(5.1) \quad \mathbf{u} = -\mathbf{f}(B'\dot{\mathbf{R}}),$$

where $\mathbf{f} : U \mapsto U$ is continuous, monotone as a graph, and satisfies $\mathbf{f}(0) = 0$. The closed-loop system is then

$$(5.2) \quad \ddot{\mathbf{R}} + B\mathbf{f}(B'\dot{\mathbf{R}}) + A\mathbf{R} = 0.$$

It follows immediately from (5.2) that, at least formally,

$$(5.3) \quad \dot{Q}(t) = \frac{1}{2} \frac{d}{dt} [\|\dot{\mathbf{R}}(t)\|_H^2 + \|\mathbf{R}(t)\|_V^2] = -(\mathbf{f}(B'\dot{\mathbf{R}}), B'\dot{\mathbf{R}}(t))_U \leq 0.$$

We are interested in determining those configurations for which a rate of decay for $Q(t)$ exists and, for such networks, in specifying the decay rate in terms of properties of the function \mathbf{f} .

It is well known that there is a close connection between exact controllability of the open-loop controlled system and the existence of a decay rate for the closed-loop system (5.2), so it should not be surprising that we can establish a decay rate only in those cases where exact controllability can be proved. In other situations, such as those considered in §4.2, we cannot obtain uniform decay estimates and, in fact, such estimates are generally not possible. To illustrate this last point, consider the linear feedback system

$$\ddot{\mathbf{R}}(t) + A\mathbf{R}(t) = B\mathbf{u}(t), \quad \mathbf{u}(t) = -B'\dot{\mathbf{R}}(t)$$

or, equivalently (cf. (4.11)),

$$(5.4) \quad \dot{\Phi}(t) = \mathcal{A}\Phi(t), \quad \mathcal{A} = \begin{pmatrix} 0 & I \\ -A & -BB' \end{pmatrix},$$

where

$$\text{Dom}(\mathcal{A}) = \{(\mathbf{R}^0, \mathbf{R}^1) | \mathbf{R}^0 \in V, \mathbf{R}^1 \in V, A\mathbf{R}^0 + B B' \mathbf{R}^1 \in H\}.$$

It is standard theory that solutions of (5.4) are given by a contraction semigroup on $V \times H$. From (5.3) we have

$$(5.5) \quad \|e^{t\mathcal{A}}\Phi^0\|_{V \times H} = \|\Phi^0\|_{V \times H} - \int_0^t \|B'\Phi(s)\|_U^2 ds.$$

Let

$$\mathcal{W} = \{\Phi^0 \in V \times H | e^{t\mathcal{A}}\Phi^0 \rightarrow 0 \text{ weakly in } V \times H \text{ as } t \rightarrow \infty\}.$$

According to a decomposition theorem due to Foguel [2], $e^{t\mathcal{A}}\mathcal{W} \subset \mathcal{W}$ for every $t \geq 0$ and $e^{t\mathcal{A}}$ is reduced to a unitary group on \mathcal{W}^\perp . By (5.5), \mathcal{W}^\perp is characterized by those Φ^0 for which $B'\Phi(s) = 0$, $s \geq 0$. This means that $\mathcal{W} = V \times H$ if and only if

$$(5.6) \quad \ddot{\mathbf{R}}(t) + A\mathbf{R}(t) = 0, \quad B'\dot{\mathbf{R}}(t) = 0, \quad t \geq 0,$$

implies $\mathbf{R}(t) \equiv 0$. However, if \mathbf{R} satisfies (5.6) then $\dot{\mathbf{R}}$ satisfies (4.16) of §4.2. The examples of that section demonstrate that for beam networks having more than one clamped node or containing a closed loop, (5.6) may admit nontrivial time periodic solutions. Such configurations therefore may not be even weakly stable under the linear feedback $\mathbf{u} = -B'\dot{\mathbf{R}}$.

On the other hand, networks that are exactly controllable to $V \times H$ are uniformly asymptotically stable in $V \times H$ under *some* linear feedback $\mathbf{u} = C\mathbf{R} + D\dot{\mathbf{R}}$, where $C \in \mathcal{L}(V, U)$ and $D \in \mathcal{L}(H, U)$ (and conversely). This may be proved via the linear quadratic regulator (LQR) framework

$$J := \min_{\mathbf{u} \in \mathcal{U}} \left(\int_0^\infty \|\Phi(t)\|_{V \times H}^2 dt + C^2 \int_0^\infty \|\mathbf{u}(t)\|_U^2 dt \right).$$

The existence of a control for which $\Phi(t) = 0$ for $t > T$ for some $T > 0$ assures that J is finite. The minimum is achieved at a unique $\mathbf{u}_0 \in \mathcal{U}$ which is given as a feedback $\mathbf{u}_0(t) = \mathcal{R}\Phi(t)$ for an appropriate Riccati operator $\mathcal{R} \in \mathcal{L}(V \times H, U)$ which, in turn, is determined through a certain algebraic operator equation. We will not pursue further the LQR approach to stabilization of beam networks in this paper. The reader is referred to the comprehensive work of Lasiecka and Triggiani [5] on the subject of Riccati equation and LQR problems for boundary/point control problems for partial differential equations.

Instead, we shall consider feedback controls of the form (5.1) in situations where it is known that the network is exactly controllable. More specifically, we consider only situations where the nonzero components of B are in the free simple nodes, in analogy with §4.1. The control space is then $\mathcal{U} = L^2(0, T, U)$ where U is defined in (4.6), and we may write

$$\mathbf{u} = - \bigotimes_{i=1}^p \mathbf{f}_i(B'\dot{\mathbf{R}}) = - \bigotimes_{i=1}^p \mathbf{f}_i(\dot{\mathbf{R}}_1(N_1), \dots, \dot{\mathbf{R}}_p(N_p)),$$

where $\mathbf{f}_i = f_{1i}\mathbf{e}_i + f_{2i}\mathbf{e}_i^\perp + f_{3i}\mathbf{n}$. In terms of components, the closed-loop system is described by (4.1)–(4.5) with $f_{ij} = f_{ij}(\dot{\mathbf{R}}_1(N_1, t), \dots, \dot{\mathbf{R}}_p(N_p, t))$ and with $\mathbf{F}_k = 0$.

The following result shows that the closed-loop system is well posed in $V \times H$.

PROPOSITION 5.1. *Let \mathcal{A} be the (nonlinear) operator in $V \times H$ defined by*

$$\mathcal{A} = \begin{pmatrix} 0 & I \\ -A & -B\mathbf{f}(B'\cdot) \end{pmatrix},$$

$$\text{Dom}(\mathcal{A}) = \{(\mathbf{R}^0, \mathbf{R}^1) \mid \mathbf{R}^0 \in V, \mathbf{R}^1 \in V, A\mathbf{R}^0 + B\mathbf{f}(B'\mathbf{R}^1) \in H\}.$$

Then \mathcal{A} is densely defined and maximal dissipative.

Proof. The monotonicity of $-\mathcal{A}$ follows immediately from that of \mathbf{f} . The statement $\text{Rg}(\lambda I - \mathcal{A}) = V \times H$ is equivalent to $\text{Rg}(A + \lambda B\mathbf{f}(B'\cdot) + \lambda^2 I) = V'$. The latter follows from the coercivity of A , the continuity and monotonicity of \mathbf{f} and $\mathbf{f}(0) = 0$. Therefore, $-\mathcal{A}$ is maximal monotone. The density of $\text{Dom}(\mathcal{A})$ in $V \times H$ follows by noting that

$$(5.7) \quad \text{Dom}(\mathcal{A}) \supset D_A \times V_0,$$

where

$$V_0 = \{\mathbf{R} \in V : R_i(N_i) = 0, i = 1, \dots, p\}.$$

Then $D_A \times V_0$ is dense in $V \times H$ and $\mathbf{f}(B'\mathbf{R}^1) = 0$ if $\mathbf{R}^1 \in V_0$, hence (5.7) holds. \square

It follows from Proposition 5.1 that given initial data $(\mathbf{R}^0, \mathbf{R}^1) \in V \times H$, (5.2) has a unique solution with $(\mathbf{R}, \dot{\mathbf{R}}) \in C([0, T]; V \times H)$ for every $T > 0$. We refer to such a solution as a finite energy solution. To study the behavior of such solutions as $t \rightarrow \infty$, we assume that the network is such that the following a priori estimate holds: for each classical solution of (4.1),

$$(5.8) \quad \int_0^T \mathcal{Q}(t) dt \leq C \left[\sup_{0 \leq t \leq T} \mathcal{Q}(t) + \sum_{i=1}^p \int_0^T Q_i(N_i, t) dt \right]$$

for some constant C independent of T .

Remark 5.1. The estimate (5.8) follows from Proposition 3.1 when the inequality (3.18) is satisfied (cf. Corollary 3.2). Sufficient conditions for (3.18) to hold were discussed in §3.1. In particular, (5.8) is valid for the networks considered in Examples 4.1 and 4.2.

When (5.8) is valid for classical solutions of (4.1) it will hold a fortiori for finite energy solutions of the closed-loop system (5.2). In fact, for such solution we have $\mathcal{Q}(t) \leq \mathcal{Q}(0)$ so that the estimate (3.13) holds. Therefore, the right-hand side of (5.8) is finite for finite energy solutions of (5.2). Moreover, from (5.3) we have

$$\sup_{0 \leq t \leq T} \mathcal{Q}(t) = \mathcal{Q}(0) = \mathcal{Q}(T) + \int_0^T (\mathbf{f}(B'\dot{\mathbf{R}}(t)), B'\dot{\mathbf{R}}(t))_U dt,$$

so that from (5.8) we obtain the estimate

$$(5.9) \quad \int_0^T \mathcal{Q}(t) dt \leq C \left[\mathcal{Q}(T) + \int_0^T (\mathbf{f}(B'\dot{\mathbf{R}}(t)), B'\dot{\mathbf{R}}(t))_U dt + \sum_{i=1}^p \int_0^T Q_i(N_i, t) dt \right]$$

for finite energy solutions of (5.2).

Decay estimates for (5.2) will be deduced from (5.9). To do so, we shall employ a proof technique introduced by Lasiecka and Tartarù [4], which they used to obtain decay rates for solutions of a class of nonlinear, dissipative wave equations. We make the following assumptions regarding \mathbf{f} .

(H1) $\mathbf{f} : U \mapsto U$ is continuous, monotone as a graph and $\mathbf{f}(0) = 0$.

(H2) For all $\mathbf{u} \in U$ with $|\mathbf{u}| \geq 1$,

$$M_1|\mathbf{u}|^2 \leq \mathbf{u} \cdot \mathbf{f}(\mathbf{u}), \quad |\mathbf{f}(\mathbf{u})| \leq M_2|\mathbf{u}|,$$

where $M_1 > 0$.

(H3) There is a concave, strictly increasing continuous function $g : \mathbb{R}^+ \mapsto \mathbb{R}$ with $g(0) = 0$ such that for all $\mathbf{u} \in U$ with $|\mathbf{u}| \leq 1$,

$$|\mathbf{u}|^{2\alpha} + |\mathbf{f}(\mathbf{u})|^2 \leq g(\mathbf{u} \cdot \mathbf{f}(\mathbf{u}))$$

for some $\alpha \in (0, 1]$.

We set

$$g_T(\xi) = g\left(\frac{\xi}{T}\right), \quad \tilde{h}_T(\xi) = (I + g_T)^{-1}(\xi/\tilde{C}), \quad \xi \geq 0, \quad T > 0,$$

where \tilde{C} is a positive constant to be specified later, and

$$h_T(\xi) = \xi - (I + \tilde{h}_T)^{-1}(\xi), \quad \xi \geq 0.$$

Note that h_T is positive and strictly increasing since \tilde{h}_T has these properties. Define $S(t)\eta$ to be the solution of the nonlinear differential equation

$$(5.10) \quad \dot{X}(t) + h_T(X(t)) = 0, \quad X(0) = \eta > 0.$$

Of course, S depends also on T . Since h_T is increasing, $S(t)$, $t \geq 0$, is a (nonlinear) contraction semigroup on \mathbb{R}^+ . The main result of this section is as follows.

THEOREM 5.2. *Assume that \mathbf{f} satisfies (H1)–(H3). Let $(\mathbf{R}, \dot{\mathbf{R}})$ be a finite energy solution of (5.2). Then there is a $T > 0$ such that*

$$\mathcal{Q}(t) \leq S(t/T - 1)\mathcal{Q}(0), \quad t \geq T,$$

where $S(t)\eta$ is the solution of (5.10).

COROLLARY 5.3. *Let \mathbf{f} satisfy (H1), (H2) and*

$$(5.11) \quad \mathbf{u} \cdot \mathbf{f}(\mathbf{u}) \geq c_0|\mathbf{u}|^{q+1}, \quad |\mathbf{f}(\mathbf{u})| \leq C_0|\mathbf{u}|^\alpha$$

for all $\mathbf{u} \in U$ with $|\mathbf{u}| \leq 1$, where $c_0 > 0$, $0 < \alpha \leq 1$, and $q \geq \alpha$. Then as $t \rightarrow \infty$,

$$\mathcal{Q}(t) = O(e^{-\omega t}) \quad \text{if } p = \alpha = 1,$$

where $\omega > 0$;

$$\mathcal{Q}(t) = O(t^{-2\alpha/(q+1-2\alpha)}) \quad \text{if } q+1 > 2\alpha.$$

Proof of Corollary 5.3. We first exhibit a function g satisfying (H3). For $|\mathbf{u}| \leq 1$ we have from (5.11)

$$\begin{aligned} |\mathbf{u}|^{2\alpha} + |\mathbf{f}(\mathbf{u})|^2 &\leq c_0^{-2\alpha/(q+1)} (\mathbf{u} \cdot \mathbf{f}(\mathbf{u}))^{2\alpha/(q+1)} + C_0^2 |\mathbf{u}|^{2\alpha} \\ &\leq c_0^{-2\alpha/(q+1)} (1 + C_0^2) (\mathbf{u} \cdot \mathbf{f}(\mathbf{u}))^{2\alpha/(q+1)}. \end{aligned}$$

Therefore (H3) is satisfied if we choose

$$g(\xi) = c_0^{-2\alpha/(q+1)} (1 + C_0^2) \xi^{2\alpha/(q+1)}.$$

Case (i). $q = \alpha = 1$. Then

$$g(\xi) = \frac{1 + C_0^2}{c_0} \xi, \quad \tilde{h}_T(\xi) = \frac{\xi}{\tilde{C}(1 + C_T)}, \quad h_T(\xi) = \frac{\xi}{1 + \tilde{C}(1 + C_T)} := \mu\xi,$$

where $C_T = (1 + C_0^2)/(c_0 T)$. Thus $S(t)\eta = e^{-\mu t}\eta$.

Case (ii). $q + 1 > 2\alpha$. We write

$$g(\xi) = \frac{1 + C_0^2}{c_0^\beta} \xi^\beta, \quad \beta = \frac{2\alpha}{q+1} < 1.$$

The function \tilde{h}_T is determined through

$$\tilde{h}_T(\tilde{C}(\xi + C_T^\beta \xi^\beta)) = \xi, \quad C_T = \frac{(1 + C_0^2)^{1/\beta}}{c_0 T}.$$

Thus asymptotically we have

$$\tilde{h}_T(\xi) \sim \frac{\xi^{1/\beta}}{C_T \tilde{C}^{1/\beta}}, \quad (\xi \rightarrow 0).$$

Furthermore, h_T satisfies

$$h_T(\xi + \tilde{h}_T(\xi)) = \tilde{h}_T(\xi),$$

so that $h_T(\xi)$ must have the same asymptotic behavior as $\tilde{h}_T(\xi)$ as $\xi \rightarrow 0$. If we define

$$H(\xi) = \int_\xi^\infty \frac{dy}{h_T(y)}, \quad \xi > 0,$$

then H is a decreasing function and $H(\eta) = 0$, $H(0+) = +\infty$. Thus \mathbb{R}^+ is in the range of H and the solution of (5.10) is given by

$$X(t) = H^{-1}(t), \quad t \geq 0.$$

Since $H(0+) = +\infty$,

$$\lim_{t \rightarrow \infty} X(t) = \lim_{t \rightarrow \infty} H^{-1}(t) = 0.$$

Let $\varepsilon > 0$, $\varepsilon < 1$. There exists $\delta(\varepsilon) > 0$ such that if $0 < \xi < \delta$,

$$|h_T(\xi) - \omega \xi^{1/\beta}| < \varepsilon \omega \xi^{1/\beta}, \quad \omega = \frac{1}{C_T \tilde{C}^{1/\beta}}.$$

Also, there exists $t_0(\varepsilon) > 0$ such that $t \geq t_0$ implies $0 < X(t) < \delta$. Therefore, if $t \geq t_0$ we have

$$-h_T(X(t)) \leq \omega(\varepsilon - 1)(X(t))^{1/\beta},$$

hence

$$\dot{X}(t) + \omega(1 - \varepsilon)(X(t))^{1/\beta} \leq 0, \quad t \geq t_0.$$

It follows that

$$X(t) = O(t^{\beta/(\beta-1)}) \quad \text{as } t \rightarrow \infty. \quad \square$$

Proof of Theorem 5.2. Since $\mathcal{Q}(t)$ is nonincreasing we have

$$\int_0^T \mathcal{Q}(t) dt \geq T\mathcal{Q}(T).$$

Substitution of this estimate into (5.9) yields, for $T > C$,

$$(5.12) \quad \mathcal{Q}(T) \leq \frac{C}{T-C} \left[\int_0^T (\mathbf{f}(B'\dot{\mathbf{R}}(t)), B'\dot{\mathbf{R}}(t))_U dt + \sum_{i=1}^p \int_0^T Q_i(N_i, t) dt \right].$$

To simplify the notation, we write

$$\mathbf{v} = B'\dot{\mathbf{R}}, \quad (\mathbf{f}(B'\dot{\mathbf{R}}), B'\dot{\mathbf{R}})_U = \mathbf{v} \cdot \mathbf{f}(\mathbf{v}).$$

We have

$$(5.13) \quad \begin{aligned} \sum_{i=1}^p \int_0^T Q_i(N_i, t) dt &= \sum_{i=1}^p \int_0^T [\rho_i |\dot{\mathbf{r}}_i|^2 + I_{\rho_i} \dot{\psi}_i^2 \\ &\quad + E_i A_i u_i'^2 + E_i I_i \psi_i'^2 + K_i (\psi_i + w_i')^2] (N_i, t) dt \\ &\leq C_1 \sum_{i=1}^p \int_0^T [|\dot{\mathbf{R}}_i(N_i, t)|^2 + |\mathbf{f}_i(\mathbf{v}(t))|^2] dt \\ &\leq C_1 \int_0^T [|\mathbf{v}(t)|^2 + |\mathbf{f}(\mathbf{v}(t))|^2] dt \end{aligned}$$

for some constant C_1 . Use of (5.13) in (5.12) gives, for $T > C$,

$$(5.14) \quad \mathcal{Q}(T) \leq \frac{C}{T-C} \left(\int_0^T \mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t)) dt + C_1 \int_0^T [|\mathbf{v}(t)|^2 + |\mathbf{f}(\mathbf{v}(t))|^2] dt \right).$$

Set

$$J_1 = \{t \in [0, T] : |\mathbf{v}(t)| \geq 1\}, \quad J_2 = [0, T] - J_1.$$

Hypothesis (H2) implies

$$\int_{J_1} [|\mathbf{v}(t)|^2 + |\mathbf{f}(\mathbf{v}(t))|^2] dt \leq M_1^{-2}(1 + M_2^2) \int_0^T \mathbf{v} \cdot \mathbf{f}(\mathbf{v}(t)) dt.$$

In addition, with hypotheses (H1) and (H3) we obtain

$$\begin{aligned} \int_{J_2} [|\mathbf{v}(t)|^2 + |\mathbf{f}(\mathbf{v}(t))|^2] dt &\leq \int_{J_2} [|\mathbf{v}(t)|^{2\alpha} + |\mathbf{f}(\mathbf{v}(t))|^2] dt \\ &\leq \int_{J_2} g(\mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t))) dt \\ &\leq \int_0^T g(\mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t))) dt. \end{aligned}$$

We apply Jensen's inequality [8, p. 359] to the last integral to the effect that

$$\begin{aligned} \int_0^T g(\mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t))) dt &\leq Tg \left(\frac{1}{T} \int_0^T \mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t)) dt \right) \\ &\leq Tg_T \left(\int_0^T \mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t)) dt \right). \end{aligned}$$

Therefore

$$(5.15) \quad \begin{aligned} \int_0^T [|\mathbf{v}(t)|^2 + |\mathbf{f}(\mathbf{v}(t))|^2] dt &\leq M_1^{-2}(1 + M_2^2) \int_0^T \mathbf{v} \cdot \mathbf{f}(\mathbf{v}(t)) dt \\ &\quad + Tg_T \left(\int_0^T \mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t)) dt \right). \end{aligned}$$

Insertion of (5.15) into (5.14) yields, for some constant C_2 , the estimate

$$(5.16) \quad \begin{aligned} \mathcal{Q}(T) &\leq \frac{C_2 T}{T - C} \left[\int_0^T \mathbf{v} \cdot \mathbf{f}(\mathbf{v}(t)) dt + g_T \left(\int_0^T \mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t)) dt \right) \right] \\ &\leq \tilde{C} \left[\int_0^T \mathbf{v} \cdot \mathbf{f}(\mathbf{v}(t)) dt + g_T \left(\int_0^T \mathbf{v}(t) \cdot \mathbf{f}(\mathbf{v}(t)) dt \right) \right] \end{aligned}$$

for $T \geq 2C$, since $T/(T - C)$ is decreasing for $T > C$, where $\tilde{C} = 2C_2$. It follows from (5.16) that

$$\tilde{h}_T(\mathcal{Q}(T)) \leq \int_0^T \mathbf{v} \cdot \mathbf{f}(\mathbf{v}(t)) dt = \mathcal{Q}(0) - \mathcal{Q}(T),$$

that is,

$$(5.17) \quad \tilde{h}_T(\mathcal{Q}(T)) + \mathcal{Q}(T) \leq \mathcal{Q}(0), \quad T \geq T_0 = 2C.$$

Now fix $T \geq T_0$. Instead of the interval $[0, T]$ we could just as well work on the interval $[mT, (m+1)T]$, $m = 1, 2, \dots$. Then (5.17) would read

$$(5.18) \quad \tilde{h}_T(\mathcal{Q}((m+1)T)) + \mathcal{Q}((m+1)T) \leq \mathcal{Q}(mT),$$

with the same function \tilde{h}_T as in (5.17). We now apply the following lemma from [4].

LEMMA 5.4 (see [4]). *Let \tilde{h} be a positive, increasing function such that $\tilde{h}(0) = 0$, and set $h(\xi) = \xi - (I + \tilde{h})^{-1}(\xi)$. Let $\{s_m\}_{m=0}^\infty$ be a sequence of positive numbers such that*

$$\tilde{h}(s_{m+1}) + s_{m+1} \leq s_m, \quad m \geq 0.$$

Then $s_m \leq S(m)s_0$ where $S(t)\eta$ is the solution of

$$\dot{X}(t) + h(X(t)) = 0, \quad X(0) = \eta \geq 0.$$

It follows from (5.18) and Lemma 5.4, applied to the sequence $s_m = Q(mT)$, that

$$Q(mT) \leq S(m)Q(0), \quad m = 0, 1, \dots$$

For any $t > 0$ we may write $t = mT + \tau$ for some integer $m \geq 0$ and $\tau \in [0, T)$. Since both $Q(t)$ and $S(t)\eta$ are nonincreasing, we have

$$Q(t) \leq Q(mT) \leq S((t - \tau)/T)Q(0) \leq S(t/T - 1)Q(0), \quad t \geq T. \quad \square$$

REFERENCES

- [1] G. CHEN, M. C. DELFOUR, A. M. KRALL, AND G. PAYRE, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.
- [2] S. R. FOGUEL, *Powers of a contraction in Hilbert space*, Pacific J. Math., 13 (1963), pp. 551–562.
- [3] J. E. LAGNESE, G. LEUGERING, AND E. J. P. G. SCHMIDT, *Modelling of dynamic networks of thin thermoelastic beams*, J. Math. Methods in Appl. Sci., (1991), to appear.
- [4] I. LASIECKA AND D. TATARU, *Uniform boundary stabilization of semilinear wave equations with nonlinear boundary conditions*, J. Differential Integral Equations, 1991, to appear.
- [5] I. LASIECKA AND R. TRIGGIANI, *Differential and algebraic Riccati equations with applications to boundary/point control problems: continuous theory and approximation theory*, Lecture Notes in Control Inform. Sci., Vol. 164, Springer-Verlag, Berlin, 1991.
- [6] G. LEUGERING AND E. J. P. G. SCHMIDT, *On the control of networks of vibrating strings and beams*, in Proc. 28th IEEE Conference on Decision and Control, 1989, pp. 2287–2290.
- [7] J. L. LIONS, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués, Tome I*, Collection RMA, Vol. 8. Masson, Paris, 1988.
- [8] J. F. RANDOLPH, *Basic Real and Abstract Analysis*, Academic Press, New York, 1968.
- [9] E. J. P. G. SCHMIDT, *On the modelling and exact controllability of networks of vibrating strings*, SIAM J. Control Optim., 30 (1992), pp. 229–245.

THE STANDARD H_∞ PROBLEM AND THE MAXIMUM PRINCIPLE: THE GENERAL LINEAR CASE*

GILEAD TADMOR†

Abstract. The classical, relatively simple ideas of linear quadratic (LQ) optimization are used in a time domain treatment of the *standard H_∞ problem*. Given the power of time domain analysis, the problem can be solved in the general framework of linear time varying, possibly infinite-dimensional, finite as well as infinite horizon systems, under no structural restrictions (such as block dimensions, zero blocks in the D operator, etc.). In the spirit of recent finite-dimensional linear time invariant (LTI) results, the solution is given in terms of two coupled Riccati equations; it includes a criterion for suboptimality and a parametrization of all suboptimal compensators. Results pertinent to LTI, periodic, and asymptotic systems are obtained as corollaries.

Key words. the standard H_∞ problem, time varying and distributed systems, LQ variational methods, differential min-max games

AMS subject classifications. 49A45, 49C05, 49C10, 90D25, 93C05, 93C35, 93C50, 93C75

1. Introduction. The celebrated H_∞ control theory has evolved in the 1980s as a *frequency domain* methodology par excellence. The theory's quintessential motivating examples—the *weighted sensitivity minimization* and the *model matching* problems—were cast in *frequency domain* terms; they have been analyzed and solved by sophisticated *frequency domain* techniques, relying mostly on factorization theory and operator interpolation (cf. [10] for an overview and references).

Yet it has recently been observed that important H_∞ results (e.g., in [3], [4], [9], [12], [13], [15], [16], [19], [22], [38], and [40]) bear considerable resemblance to, by now classical, linear quadratic Gaussian (LQG) observations. And while optimal control problems have been treated from algebraic perspectives, the inherent logic of the field stems from studies of system dynamics, i.e., from *time domain* analysis. Dynamic programming and the maximum principle are perhaps the two most fundamental observations in that respect.

It has also been observed [10] that H_∞ problems (and for that matter, a wide range of worst-case design issues) can be cast in the transform-invariant terms of input-output (I/O) operator norms. That is, these problems are equally meaningful over the *time domain* as over the *frequency domain*.

Motivated by these observations, we develop here an interpretation of the generic *standard problem* as a competition between disturbances and controls, with a quadratic cost objective. We are thus able to effectively invoke the powerful yet relatively simple LQ maximum principle as a main tool to obtain our version of the recent *two Riccati equations* result. It includes a suboptimality criterion and a parametrization of all suboptimal solutions. We do so in a general setting of linear systems, making no structural restrictions and allowing any time horizon, time varying systems, and infinite dimensionality.

A summary of the present discussion and some follow-ups are given in [36] and [37]. This note extends and complements our preliminary investigation of the same issue in [30] and [31] (where our LQ approach was first presented) and of I/O norms in linear systems in [32]. Its main results (Theorems I and II) are intended as a

* Received by the editors May 29, 1989; accepted for publication (in revised form) November 22, 1991.

† Department of Electrical and Computer Engineering, 409 Dana Research Building, Northeastern University, Boston, Massachusetts 02115.

theoretical backdrop to some further studies, both theoretical and applied, in robust control. A hint on examples of subclasses of linear time varying (LTV) systems where they can be feasibly implemented is given in the corollary, while some further developments stemming from them are presented in [33]–[35]. It is noted that similar and related approaches have been taken recently by several other authors (e.g., in [2], [17], [25]–[27], [29], and [39]).

In dealing with the present level of generality it seems essential to stick to a purely dynamic line of arguments, excluding many shortcuts we usually make in the LTI case (e.g., using linear algebraic techniques) or even differentiation (as the finite-dimensional case allows). Most of those arguments are fairly simple, drawn from a standard inventory of LQ optimization and linear dynamic systems tools. Yet their accumulation is felt in this note's length and notation burden; and while some of the resulting encumbrance is certainly due to the author's shortcomings, it seems that most of it is an inevitable consequence of the undertaking. We shall try to ease readability by occasional footnotes.

The paper is organized as follows: Preliminaries and the statement of main results are given in § 2. Proofs are given in § 3. For the sake of clarity, these results and proofs are made under certain simplifying assumptions on the system's structure. In § 4 we discuss the general case, where the simplifying structural assumptions are dropped.

2. Preliminaries. We consider a linear system \mathcal{S} of the form

$$(1) \quad \begin{aligned} \dot{x} &= Ax + B_1 w + B_2 u, \\ z &= C_1 x + D_{11} w + D_{12} u, \\ y &= C_2 x + D_{21} w + D_{22} u \end{aligned}$$

over a time interval (t_0, t_1) , $-\infty \leq t_0 < t_1 \leq +\infty$. In this setting x , u , w , y , and z are interpreted as the state, control, disturbance, observation, and the output signals, respectively. These signals take values in the Hilbert spaces \mathbf{X} , \mathbf{U} , \mathbf{W} , \mathbf{Y} , and \mathbf{Z} . The coefficients A , B_i , C_i , and D_{ij} are allowed to vary in time (e.g., $A = A(t)$). The input and output coefficients B_i , C_i , and D_{ij} are assumed to be L_∞ bounded-operator-valued functions. The operator A is assumed to generate a uniformly exponentially bounded evolution system $\Phi(t, s)$ on \mathbf{X} . That is, there exist real constants α and β such that¹

$$(2) \quad \|\Phi(t, s)\| \leq \alpha e^{\beta(t-s)}$$

for $t \geq s$.

Note. The reader who prefers to focus on the finite-dimensional case may substitute “transition matrix” for “evolution system” and assume that A is an L_∞ matrix-valued function. In the infinite-dimensional LTI case, “ c_0 -semigroup” should substitute for “evolution system” and exponential boundedness is guaranteed by definition. Details and precise definitions pertinent to evolution systems and their generators, as well as examples of conditions that assure that an operator generates an evolution system, can be found in [5]–[8], [11], [14], and [20].

For the benefit of those who wish to focus only on the finite-dimensional case, we chose to stick to the differential equation formalism, as in (1), and use the terms “generator” (and thereby “generates,” “generated,” etc.) also in reference to bounded perturbations of true generators. *It is stressed that these conventions do not adhere to*

¹ We use the same notation $\|\cdot\|$ in reference to the norm of a Hilbert space vector, a time function, and an operator's induced norm. The meaning will be clear from the context. When in doubt, we use subscript in reference to the underlying space, e.g. $\|\xi\|_X$. We also use subscript to denote the weight of a weighted Hilbert space norm, e.g. $\|\varphi\|_\Omega^2 = \langle \varphi, \Omega \varphi \rangle$. The reader will be alerted of the latter notation when it is used.

standard definitions and are made solely for the purpose of simplicity in notation and terminology. A brief discussion of the appropriate interpretation of (1) and of our (ab)use of the term “generator” and its derivatives, is provided in the remark on infinite dimensionality and in Lemmas 2.1 and 2.2 at the end of this section. Those readers who are interested in the distributed case may wish to consult that remark at this point.

Restrictions of (1) to subintervals $[t, t_1)$, $t \in (t_0, t_1)$ will be considered. Associated with each such restriction are the linear (bounded, in the finite horizon case) mappings from the initial state and the inputs u and w to the state, observation and output trajectories. Those will be denoted, respectively, by

$$\chi(t): (x(t), u, w) \rightarrow x: X \times L_2(t, t_1) \times L_2(t, t_1) \rightarrow \times L_2(t, t_1),$$

$$\mathcal{Y}(t): (x(t), u, w) \rightarrow y: X \times L_2(t, t_1) \times L_2(t, t_1) \rightarrow \times L_2(t, t_1),$$

and

$$\mathcal{X}(t): (x(t), u, w) \rightarrow z: X \times L_2(t, t_1) \times L_2(t, t_1) \rightarrow \times L_2(t, t_1).$$

As the discussion unfolds we introduce and use variants of these notations, in reference to the mappings associated with feedback control or the result of certain optimization problems. Those variants will be distinguished by super- and subscript notation, as will be explained in due time.

An *admissible feedback operator* $u = \mathcal{K}y$ in (1) is defined in terms of a linear time varying system

$$(3) \quad \dot{p} = Mp + Ny, \quad u = Qp + Ry,$$

where p takes values in a Hilbert space \mathbf{P} and where the assumptions on the coefficients in (3) are analogous to those made in (1). The closed-loop system, depicted in Fig. 1, should be well defined under our feedback. That is guaranteed by the condition that $I - RD_{22}$ (or equivalently, $I - D_{22}R$) be invertible. In the present time-varying setting, we require that admissible compensators maintain uniform invertibility; i.e., $(I - RD_{22})^{-1}$, $(I - D_{22}R)^{-1} \in L_\infty$. It is important to remember that since the feed-through component R is invariant under the choice of a particular realization of a compensator, so is the uniform invertibility condition.

The free motion in the closed-loop system is governed by this next operator:

$$(4) \quad \mathcal{A} = \begin{bmatrix} A + B_2(I - RD_{22})^{-1}RC_2 & B_2(I - RD_{22})^{-1}Q \\ N(I - D_{22}R)^{-1}C_2 & M + ND_{22}(I - RD_{22})^{-1}Q \end{bmatrix}.$$

Our admissible feedback is *internally stabilizing* if the operator \mathcal{A} is exponentially stable; i.e., such that “ β ” in the closed-loop counterpart of (2) is negative.

It is noted that in the case of a static feedback, $u = Ry$ the state space \mathbf{P} becomes trivial and (4) reduces to

$$(5) \quad \mathcal{A} = A + B_2(I - RD_{22})^{-1}RC_2.$$

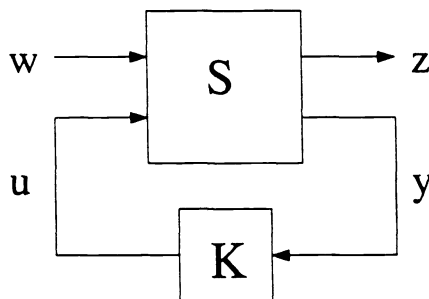


FIG. 1. The closed-loop system.

Given an internally stabilizing feedback \mathcal{K} we denote by $\mathcal{T}_{\mathcal{K}}(s)$ the closed-loop mapping $w \rightarrow z = \mathcal{Z}(s)(0, u = \mathcal{K}y, w): L_2(s, t_1) \rightarrow L_2(s, t_1)$, $s \in (t_0, t_1)$. These operators are bounded, uniformly for all choice of s . Obviously, $\|\mathcal{T}_{\mathcal{K}}(t_0)\| = \max_s \|\mathcal{T}_{\mathcal{K}}(s)\|$. The optimal value in (1) is

$$(6) \quad \gamma_0 = \inf \{ \|\mathcal{T}_{\mathcal{K}}(t_0)\| : \mathcal{K} \text{ is internally stabilizing} \}.$$

In the context of the *standard problem* we want (i) to characterize *suboptimal values* $\gamma > \gamma_0$, and (ii) given a suboptimal γ , to parametrize all internally stabilizing compensators that assure $\|\mathcal{T}_{\mathcal{K}}(t_0)\| < \gamma$.

The following assumptions on the structure of \mathcal{S} result in considerable notational simplification of our main results and their proofs, and will be made throughout the remainder of this section and in § 3. In § 4 we will discuss the mathematically simple but notationally awkward steps that may be taken in removing all these assumptions.

$$(7) \quad \begin{aligned} D_{11} &= 0, & D'_{12}C_1 &= 0, & D'_{12}D_{12} &= I, & C'_1C_1 &\geq \varepsilon I, \\ D_{22} &= 0, & D_{21}B'_1 &= 0, & D_{21}D'_{21} &= I, & B_1B'_1 &\geq \varepsilon I. \end{aligned}$$

Subject to (7), the following are our main results.

THEOREM I. (a) $\gamma > \gamma_0$ if and only if there exist negative-definite, operator-valued functions $\Pi_1, \Pi_2 \in L_\infty(t_0, t_1)$, $(\Pi_i(t): X \rightarrow X)$ such that

(i) The operators $A_1 = A + (B_2B'_2 - (1/\gamma^2)B_1B'_1)\Pi_1$ and $A_2 = A - (1/\gamma^2)B_1B'_1\Pi_1 + \Pi_2(C'_2C_2 - (1/\gamma^2)\Pi_1B_2B'_2\Pi_1)$ generate exponentially stable evolution systems, Φ_1 and Φ_2 , respectively.

(ii) Π_1 and Π_2 satisfy these next Riccati integral equations:

$$(8) \quad \Pi_1(s)\xi = - \int_s^{t_1} \Phi'_1(t, s) \left[C'_1C_1 + \Pi_1 \left(B_2B'_2 - \frac{1}{\gamma^2} B_1B'_1 \right) \Pi_1 \right] (t) \Phi_1(t, s) \xi dt$$

and

$$(9) \quad \Pi_2(t)\xi = - \int_{t_0}^t \Phi_2(t, s) \left[B_1B'_1 + \Pi_2 \left(C'_2C_2 - \frac{1}{\gamma^2} \Pi_1B_2B'_2\Pi_1 \right) \Pi_2 \right] \cdot (s) \Phi'_2(t, s) \xi ds.$$

(b) Suppose that $\gamma > \gamma_0$. Then the following is a parametrization of all internally stabilizing compensators for (1) that assure $\|\mathcal{T}_{\mathcal{K}}(t_0)\| < \gamma$:

$$(10) \quad \begin{aligned} \dot{p} &= (A_1 + \Pi_2C'_2C_2)p + \Pi_2C'_2y - \left(I + \frac{1}{\gamma^2} \Pi_2\Pi_1 \right) B_2v, \\ q &= C_2p + y, & v &= \mathcal{K}_0q, \\ u &= -B'_2\Pi_1p + v. \end{aligned}$$

The free design parameter \mathcal{K}_0 may be any admissible compensator in (10), realized by a stable system and satisfying $\|\mathcal{K}_0\| < \gamma$.

THEOREM II. Assume state feedback is allowed (that is, substitute in (7) $C_2 = I$, $D_{21} = 0$). Then the following holds:

(c) $\gamma > \gamma_0$ if and only if Π_1 exists, as specified in part (a).

(d) Assume $\gamma > \gamma_0$. Then $u = B'_2\Pi_1x$ is an internally stabilizing compensator that assures $\|\mathcal{T}_{\mathcal{K}}(t_0)\| < \gamma$.

COROLLARY. (e) Assume that A is an L_∞ , bounded-matrix-valued function. Then, when existing, the solutions to the Riccati equations (8), (9) are absolutely continuous, and satisfy the differential counterparts of these equations:

$$(11) \quad \dot{\Pi}_1 = C'_1C_1 - \Pi_1A - A'\Pi_1 - \Pi_1 \left(B_2B'_2 - \frac{1}{\gamma^2} B_1B'_1 \right) \Pi_1$$

and

$$(12) \quad \begin{aligned} \dot{\Pi}_2 = & -B_1 B_1' + \Pi_2 \left(A - \frac{1}{\gamma^2} B_1 B_1' \Pi_1 \right)' + \left(A - \frac{1}{\gamma^2} B_1 B_1' \Pi_1 \right) \Pi_2 \\ & + \Pi_2 \left(C_2' C_2 - \frac{1}{\gamma^2} \Pi_1 B_2 B_2' \Pi_1 \right) \Pi_2, \end{aligned}$$

so that (i) if $t_1 < +\infty$ then $\Pi_1(t_1) = 0$; else $\Pi_1^{-1} \in L_\infty$; and (ii) if $t_0 > -\infty$ then $\Pi_2(t_0) = 0$; else $\Pi_2^{-1} \in L_\infty$.

(f) If the system is defined over the entire real line $((t_0, t_1) = (-\infty, +\infty))$ and its coefficients are periodic, then so are the solutions of the Riccati equations (8), (9) (or (11), (12)).

(g) If (1) is time invariant and defined over an infinite interval then Π_1 and Π_2 are constant operators and algebraic Riccati equations substitute for their differential counterparts in part (e).

(h) Suppose that $t_0 = -\infty$ and $t_1 = +\infty$, and that there exist $A^+, B_i^+, C_i^+, D_{ij}^+, A^-, B_i^-, C_i^-$, and D_{ij}^- such that the coefficients in (1) tend (in operator norm) to their “+” counterparts as $t \rightarrow +\infty$ and to their “-” counterparts as $t \rightarrow -\infty$. (e.g., $\|A(t) - A^+\| \rightarrow 0$ as $t \rightarrow +\infty$). Then $\Pi_i(t) \rightarrow \Pi_i^\pm$ as $t \rightarrow \pm\infty$, where Π_i^\pm are the solutions of the \pm limiting, time-invariant Riccati equations. In particular, if γ is strictly suboptimal in (1), so it is in the limiting systems. A γ -suboptimal, internally stabilizing compensator in the “+” system, say \mathcal{K}^+ , is internally stabilizing in (1), and it guarantees that eventually $\|\mathcal{T}_{\mathcal{K}^+}(t)\| < \gamma$.

Remark on handling infinite dimensionality. Our main purpose in this note is to adapt the relatively intuitive and simple arguments of time-domain LQ optimization, so that they could be utilized in solving the *standard H_∞ problem*, in as general a linear setup as may be reasonably desired. Since our aim is *not* the study of infinite-dimensional systems per se, we thus make the easily understood, simply stated assumption that “*A generates an exponentially bounded evolution system*,” rather than get into the detail of specific conditions on A that guarantee that that assumption holds. For such detail we refer the interested reader to the extensive literature on infinite-dimensional systems and, in particular, the citations made above. In line with the same philosophy, and to keep the presentation digestible for those who would rather focus on the finite-dimensional case, we further use certain formal simplifications.

First, as we have already noted above, we stick to the differential equation formalism, as in (1). Since mild evolutions might not be differentiable in the distributed case, all systems should then be interpreted in terms of the associated integral “variation of parameters” formulae, as done, e.g., in [11]. It is stressed that none of the developments in this paper requires state differentiation, and all of them could be made in the framework of specified integral “variation of parameters” input-output mappings.

Second, since we nonetheless use the differential equation form and the term “generator,” it is noted that even when an operator adheres to the requirements of a proper definition of a generator, the same is not necessarily true for bounded perturbations of that operator. Yet such bounded perturbations play an important role in this discussion (especially \mathcal{A} and A_i , $i = 1, 2$, above). As mentioned above, for the sake of simplicity we use the term “generator” also in reference to boundedly perturbed “true” generators. Those should be interpreted in terms of the following well-known lemma.

LEMMA 2.1 [6]. Suppose that E generates an exponentially bounded evolution system Ψ over the Hilbert space \mathbf{H} , and that F is an L_∞ bounded-operator-valued function. Then there exists a unique exponentially bounded evolution system $\tilde{\Psi}$ satisfying the following

integral equation:

$$(13) \quad \tilde{\Psi}(t, s)h = \Psi(t, s)h + \int_s^t \Psi(t, r)F(r)\tilde{\Psi}(r, s)h \, dr$$

for all $h \in \mathbf{H}$.

For the purpose of this discussion we deviate from standard terminology and refer to $\tilde{\Psi}$ as *the evolution generated by the perturbed generator $E + F$* (equivalently, we shall say that $E + F$ generates $\tilde{\Psi}$). Manipulations of the underlying differential equation model of a system, such as the introduction of feedback, can always be interpreted in terms of appropriate manipulations of the associated integral representations, in line with the lemma. The infinite-dimensional expert will be able to easily fill in the details, and thus we mostly omit them. Exceptions will be made and details provided in less obvious cases.

Outline of the proof. The usual: It easily follows from the exponential growth bound

$$(14) \quad \|\Psi(t, s)\| \leq \alpha e^{\beta(t-s)}, \quad t \geq s$$

that the successive approximations

$$(15) \quad \begin{aligned} \Psi_0 &= \Psi, \\ \Psi_{i+1}(t, s)h &= \Psi(t, s)h + \int_s^t \Psi(t, r)F(r)\Psi_i(r, s)h \, dr, \quad i = 0, 1, 2, \dots \end{aligned}$$

converge, uniformly for all h in the unit ball, and that the limit, which is a solution of (13), satisfies the growth condition

$$(16) \quad \|\tilde{\Psi}(t, s)\| \leq \alpha e^{(\beta + \alpha \|F\|_\infty)(t-s)}, \quad t \geq s;$$

hence *existence*.

Given any exponentially bounded solution, a successive substitution of the entire right-hand side of (13) for $\tilde{\Psi}$ in the integral term of that equation, shows that that solution must indeed be the limit of the successive approximations (15). Hence *uniqueness*. \square

Finally, the following result provides a very useful tool, to be used in stability analysis of distributed systems.

LEMMA 2.2 [8]. *Let E and F be as above and assume that E is exponentially stable; that is, that β in (14) is negative. Suppose also that the following inequality holds for some fixed η and all $t_0 < s < t_1$:*

$$(17) \quad \int_s^{t_1} \|\tilde{\Psi}(r, s)h\|^2 \, dr \leq \eta \|h\|^2.$$

Then the perturbed generator $E + F$ is also exponentially stable.

Note. The significance of this lemma is in either of the cases $t_0 = -\infty$ or $t_1 = +\infty$.

Proof. It suffices to focus only on the integral term on the right-hand side of (13). We first observe that that term is uniformly bounded over all choice of $t_1 > t \geq s > t_0$. Indeed,

$$(18) \quad \begin{aligned} \left\| \int_s^t \Psi(t, r)F(r)\tilde{\Psi}(r, s)h \, dr \right\|^2 &\leq \int_s^t \|\Psi(t, r)\|^2 \, dr \|F\|_\infty^2 \int_s^{t_1} \|\tilde{\Psi}(r, s)h\|^2 \, dr \\ &\leq \alpha^2 \frac{(e^{2\beta(t-s)} - 1)}{2\beta} \|F\|_\infty^2 \eta \|h\|^2 \\ &\leq \frac{\alpha^2}{2|\beta|} \|F\|_\infty^2 \eta \|h\|^2 < \infty. \end{aligned}$$

Next we show that the bound (17) can be improved to the following:

$$(19) \quad \int_p^{t_1} \|\tilde{\Psi}(r, s)h\|^2 dr \leq \eta \|h\|^2 e^{(s-p)/\eta}, \quad p \geq s.$$

Indeed, invoking (17) we see that the function

$$(20) \quad f(p) \triangleq \int_p^{t_1} \|\tilde{\Psi}(r, s)h\|^2 dr, \quad p \geq s$$

satisfies the differential and initial value inequalities

$$(21) \quad \begin{aligned} f(p) &= \int_p^{t_1} \|\tilde{\Psi}(r, s)h\|^2 dr = \int_p^{t_1} \|\tilde{\Psi}(r, p)\tilde{\Psi}(p, s)h\|^2 dr \\ &\leq \eta \|\tilde{\Psi}(p, s)h\|^2 = \eta(-\dot{f}(p)), \quad p \geq s, \end{aligned}$$

and

$$f(s) = \int_s^{t_1} \|\tilde{\Psi}(r, s)h\|^2 dr \leq \eta \|h\|^2,$$

which entails (19).

Now we split the integral term in (13) into two:

$$(22) \quad \int_s^{(t+s)/2} \Psi(t, r)F(r)\tilde{\Psi}(r, s)h dr + \int_{(t+s)/2}^t \Psi(t, r)F(r)\tilde{\Psi}(r, s)h dr.$$

In view of (18), the first term in (22) satisfies

$$(23) \quad \begin{aligned} &\left\| \int_s^{(t+s)/2} \Psi(t, r)F(r)\tilde{\Psi}(r, s)h dr \right\|^2 \\ &\leq \left\| \Psi\left(t, \frac{t+s}{2}\right) \right\|^2 \left\| \int_s^{(t+s)/2} \Psi\left(\frac{t+s}{2}, r\right)F(r)\tilde{\Psi}(r, s)h dr \right\|^2 \\ &\leq \alpha^2 e^{\beta(t-s)} \frac{\alpha^2}{2|\beta|} \|F\|_\infty^2 \eta \|h\|^2, \end{aligned}$$

where by (19), the second term obeys the bound

$$(24) \quad \begin{aligned} &\left\| \int_{(t+s)/2}^t \Psi(t, r)F(r)\tilde{\Psi}(r, s)h dr \right\|^2 \\ &\leq \int_{(t+s)/2}^t \|\Psi(t, r)\|^2 dr \|F\|_\infty^2 \int_{(t+s)/2}^{t_1} \|\tilde{\Psi}(r, s)h\|^2 dr \\ &\leq \alpha^2 \frac{(e^{2\beta(t-s)/2} - 1)}{2\beta} \|F\|_\infty^2 \eta e^{(s-t)/2\eta} \|h\|^2 \\ &\leq \frac{\alpha^2}{2|\beta|} \|F\|_\infty^2 \eta e^{(s-t)/2\eta} \|h\|^2. \end{aligned}$$

Both of these bounds indicate exponential decay, which completes the proof. \square

3. Proofs. The underlying idea is simple. It is based on the following straightforward observation.

OBSERVATION 3.0.1. $\gamma > \gamma_0$ if and only if there exist an internally stabilizing compensator \mathcal{K} and a constant $\delta > 0$ such that the following inequality holds for all w in L_2 :

$$(25) \quad \gamma^2 \|w\|^2 - \|\mathcal{T}_{\mathcal{K}}(t_0)w\|^2 \geq \delta^2 \|w\|^2.$$

With this observation in mind we define the family of quadratic cost indexes (parametrized by $s \in (t_0, t_1)$)

$$(26) \quad J(s; x(s), u, w) \triangleq \gamma^2 \|w\|_{L_2(s, t_1)}^2 - \|z\|_{L_2(s, t_1)}^2$$

where z is the system output along (s, t_1) , given the initial state $x(s)$ and the input functions u, w . We shall explore the following min-max problems:

$$(27) \quad \min_{w \in L_2} \max_{u \in L_2} J(s; \xi, u, w).$$

As stated, (27) is an open-loop problem; but following the example of other LQ optimization problems, we may well expect the solution to be given in closed loop. Therefore, we interpret the goal of the maximizing control as maintaining (25), whereas the minimizing disturbance tries to violate it.

Remark. A great deal of our efforts in what follows will be devoted to establishing stability and related boundedness properties. These parts of the discussion make sense, of course, and should be understood only in the context of systems that are defined over an infinite time interval. The finite horizon proof is considerably simpler, and we leave it to the reader to trim the redundancies for that case.

3.1. The LQ optimization problem. We start with the solution of the max part of (27). Throughout this section we assume that (1) is internally stabilizable; that is, that there exists an internally stabilizing compensator \mathcal{K} (of the form (3)) in (1). This assumption holds whenever the *optimal value* in (1), γ_0 , is finite. We then define²

$$(28) \quad c(s; \xi, u^0, w) \triangleq \|\mathcal{Z}(s)(\xi, u, w)\|^2 = \|C_1 \mathcal{X}(s)(\xi, u, w)\|^2 + \|u\|^2$$

and look for the control u^0 that satisfies

$$(29) \quad c(s; \xi, u^0, w) = c^0(s; \xi, w) \triangleq \inf_u c(s)(\xi, u, w).$$

In view of (7), $c(s)$ is a positive quadratic cost functional, and is nonsingular with respect to both state and control. That is, there holds

$$(30) \quad c(s; \xi, u, w) \geq \|u\|^2, \quad c(s; \xi, u, w) \geq \varepsilon \|\chi(s)(\xi, u, w)\|^2.$$

The solution to (29), which we bring for completeness and for reference in later parts of the proof, will follow the usual optimal control line (see, e.g., [1] and [18]): First we establish existence and uniqueness of u^0 . Then we shall characterize it in terms of a Hamilton–Jacobi boundary-value problem. Finally, the state and co-state of the latter will be related by the solution of a Riccati equation.

PROPOSITION 3.1.1. *If an internally stabilizing compensator exists in (1) then there is $\theta > 0$ such that for all $s \in (t_0, t_1)$ there holds*

$$(31) \quad c^0(s; \xi, w) < \theta^2 \|(\xi, w)\|^2.$$

Proof. Obviously, there holds

$$(32) \quad c^0(s; \xi, w) \leq \|\mathcal{Z}_{\mathcal{K}}(s)(\xi, w)\|^2 \leq \|\mathcal{Z}_{\mathcal{K}}(s)\|^2 \|(\xi, w)\|^2,$$

² The second equality in (28) is due to (7).

where \mathcal{K} is some stabilizing compensator, and where $\mathcal{L}_{\mathcal{K}}(s)(\xi, w) \triangleq \mathcal{L}(s)(\xi, u = \mathcal{K}y, w)$ defines the associated bounded, closed-loop output operator. Set $\theta = \|\mathcal{L}_{\mathcal{K}}(t_0)\|$. \square

PROPOSITION 3.1.2. *Assume (1) is internally stabilizable. Then, given s , ξ , and w , there exists a unique optimal control $u^0 = \mathcal{U}^0(s)(\xi, w)$ that attains the minimum in (29). Moreover, the operators $\mathcal{U}^0(s)$ are bounded uniformly for all choices of s .*

Proof. Given s , ξ , w , u_1 and u_2 such that all the following terms are finite, there holds

$$\begin{aligned} (7, i) \Rightarrow \|u_1 - u_2\|^2 &\leq \|\mathcal{L}(s)(0, u_1 - u_2, 0)\|^2 \\ &= 2(\|\mathcal{L}(s)(\xi, u_1, w)\|^2 + \|\mathcal{L}(s)(\xi, u_2, w)\|^2 \\ &\quad - 2\|\mathcal{L}(s)(\xi, \frac{1}{2}(u_1 + u_2), w)\|^2) \\ &\leq 2(\|\mathcal{L}(s)(\xi, u_1, w)\|^2 + \|\mathcal{L}(s)(\xi, u_2, w)\|^2 - 2c^0(s; \xi, w)). \end{aligned} \quad (33)$$

Therefore, if $\{u_\alpha\}$ is a minimizing sequence (i.e., $c^0(s; \xi, w) = \lim J(s; \xi, u_\alpha, w)$), it is a Cauchy sequence in L_2 , tending to a limit u^0 . Moreover, the limit is unique. Since $u \rightarrow c(s)(\xi, u, w)$ is a close mapping, u^0 satisfies (29). By (30) and (31) we have

$$\|\mathcal{U}^0(s)\| \leq \theta, \quad (34)$$

which completes the proof. \square

Having the proposition we introduce notation for the optimal state and output mappings $\mathcal{X}^0(x_s)(\xi, w) \triangleq \mathcal{X}(s)(\xi, u^0, w)$, $\mathcal{Z}^0(s)(\xi, w) \triangleq \mathcal{Z}(s)(\xi, u^0, w)$. It follows from (30) and (31) that these too are bounded operators, uniformly for all choice of s .

PROPOSITION 3.1.3. *Assume (1) is internally stabilizable. Then, given s , ξ , and w , as above, there exists a unique solution to the following inhomogeneous Hamilton–Jacobi boundary value problem:³*

$$\begin{aligned} \dot{x} &= Ax + B_2 B_2' e + B_1 w, & x(s) &= \xi, \\ \dot{e} &= C_1' C_1 x - A' e, & e(t_1) &= 0. \end{aligned} \quad (35)$$

Moreover, let (x, e) be the solution. Then $x = \mathcal{X}^0(s)(\xi, w)$ and $\mathcal{U}^0(s)(\xi, w) = B_2' e$. Denote $e = \mathcal{E}^0(s)(\xi, w)$. Then the linear mappings $\mathcal{E}^0(s)$ are bounded, uniformly for all choice of s .

Proof. Since we can only deal with the integrated form of (35), it will simplify matters if we assume for the moment that there exists an internally stabilizing state feedback compensator $u = Kx$ in (1), defined by an operator-valued function $K \in L_\infty$ with $K(t) : X \rightarrow U$. Indeed, it is established in [23] that dynamic feedback stabilizability (which is the current underlying hypothesis) implies the existence of a stabilizing static feedback, so this assumption causes no restriction of generality. For completeness we state and establish this fact in Lemma 3.1.6. Let

$$A_K \triangleq A + B_2 K \quad (36)$$

be the closed-loop exponentially stable generator, as in (5), and let Φ_K be the evolution system generated by A_K . Using the change of variables

$$u = Kx + v \quad (37)$$

the first and third equations in (1) read

$$\dot{x} = A_K x + B_2 v + B_1 w, \quad z = (C_1 + D_{12} K)x + D_{12} v, \quad (38)$$

³ In case $t_1 = +\infty$ the terminal constraint should be interpreted as follows: $\exists \lim_{t \rightarrow +\infty} e(t) = 0$.

which defines a stable system, and where the optimization variable is v .⁴ We denote by $\mathcal{Z}_K(s)(\xi, v, w)$ the output mapping in (38).

The following is an outline of a standard calculus of variations reasoning. Let s , ξ , and w be given and fix v^+ , $v \in L_2$. Then

$$(39) \quad \begin{aligned} \|\mathcal{Z}_K(s)(\xi, v^+ + v, w)\|^2 &= \|\mathcal{Z}_K(s)(\xi, v^+, w)\|^2 \\ &+ 2\langle \mathcal{Z}_K(s)(\xi, v^+, w), \mathcal{Z}_K(s)(0, v, 0) \rangle \\ &+ \|\mathcal{Z}_K(s)(0, v, 0)\|^2. \end{aligned}$$

Since the last term on the right-hand side of (39) is nonnegative, v^+ is optimal (i.e., it minimizes the norm of $\mathcal{Z}_K(\xi, v^+, w)$) if and only if

$$(40) \quad \langle \mathcal{Z}_K(s)(\xi, v^+, w), \mathcal{Z}_K(s)(0, v, 0) \rangle = 0$$

for all $v \in L_2$. Explicitly, this condition reads

$$(41) \quad u^+ = B_2' e^+,$$

where u^+ corresponds to v^+ via (37), and where e^+ is defined by⁵

$$(42) \quad e^+(t) = - \int_t^{t_1} \Phi_K'(s, t) (C_1' C_1 x^+ + K' u^+)(s) ds.$$

Finally, formal differentiation⁶ shows that the system comprising the first equation in (38) and equations (41) and (42) is equivalent to the Hamilton–Jacobi system (35), when restricted to L_2 trajectories.

Since there does exist an optimal control in (1), solving (29), this establishes the existence of a solution to (35).

Conversely, by the arguments above, any solution to (35) and the relation (41) define the unique optimal state and control in (29). If t_1 is finite, the co-state is completely determined by (42), which establishes uniqueness of the solution to (35). It remains to establish uniqueness in the case $t_1 = +\infty$. To that end we show that even when $t_1 = +\infty$ it is impossible to have a nonzero function e satisfying (35) and (41) with $x=0$, $u=0$ and $w=0$. The proof, as in the finite-dimensional case, relies on the assumption that (1) is stabilizable.

Indeed, suppose such $e(t)$ did exist. Then the stability of Φ_K and the assumptions $e(s) = \Phi(t, s)' e(t)$, $\lim_{t \rightarrow +\infty} e(t) = 0$, and $u = B_2' e = 0$ imply that for every $s < t$ and $\xi \in X$ there holds

$$(43) \quad \begin{aligned} 0 &= \int_s^t \langle B_2'(r) e(r), K(r) \Phi_K(r, s) \xi \rangle dr \\ &= \left\langle e(t), \int_s^t \Phi(t, r) B_2(r) K(r) \Phi_K(r, s) \xi dr \right\rangle \\ &= \langle e(t), \Phi_K(t, s) \xi \rangle - \langle e(t), \Phi(t, s) \xi \rangle \\ &= \langle e(t), \Phi_K(t, s) \xi \rangle - \langle e(s), \xi \rangle \xrightarrow{t \rightarrow +\infty} -\langle e(s), \xi \rangle \end{aligned}$$

whereby $e=0$, as claimed.

⁴ It is obvious that both u and x are L_2 functions if and only if the function v belongs to L_2 . Hence optimization over v is equivalent to optimization over u .

⁵ Our conversion of the system to a stable one, via (37), was made with the purpose that (42) be well defined. That is a usual trick. Once its role in establishing this proposition is done, (37) will be replaced by more advantageous feedback.

⁶ We recall that differentiation is generally not allowed in an infinite-dimensional setting, which is why we added the adjective “formal.” With little work the equivalence can be established in terms of the integral equation of Lemma 2.1 and the definition of boundedly perturbed generators that follows that lemma.

Now we can substitute x^0 , u^0 , and e^0 for x^+ , u^+ , and e^+ in (42), and get an expression for $\mathcal{E}^0(s)$ in terms of $\mathcal{U}^0(s)$ and of $\mathcal{X}^0(s)$. Uniform boundedness of $\mathcal{E}^0(s)$ is the result of the uniform boundedness of $\mathcal{U}^0(s)$ and of $\mathcal{X}^0(s)$, and from the exponential stability of A_K . This completes the proof. \square

Denote

$$(44) \quad \Lambda(s)\xi \triangleq (\mathcal{E}^0(s)(\xi, 0))(s).$$

By the previous proposition (in particular, as follows from (42)), these are uniformly bounded operators. Set

$$(45) \quad A_0 \triangleq A + B_2 B_2' \Lambda.$$

PROPOSITION 3.1.4. A_0 is an exponentially stable generator.

Proof. Let Φ_0 be the evolution generated by A_0 . By Proposition 3.1.3 there holds

$$(46) \quad \Phi_0(t, s)\xi = (\chi^0(s)(\xi, 0))(t).$$

Following from the L_2 uniform boundedness of $\mathcal{X}^0(s)$ there must exist η such that

$$(47) \quad \forall \xi, \quad \int_s^{t_1} \|\Phi_0(t, s)\xi\|^2 ds < \eta \|\xi\|^2$$

for all s . The proposition thus follows from Lemma 2.2, substituting A_K for E and $B_2' \Lambda - K$ for F . \square

PROPOSITION 3.1.5. $\Xi = \Lambda$ is the unique uniformly-bounded, negative-definite strong solution of the following integral Riccati equation.⁷

$$(48) \quad \Xi(s)\xi = - \int_s^{t_1} \Psi(r, s)' (C_1' C_1 + \Xi B_2 B_2' \Xi)(r) \Psi(r, s) \xi dr,$$

such that Ψ is the exponentially stable evolution generated by $A_\Xi \triangleq A + B_2 B_2' \Xi$.

Proof. Using the definition (44) for Λ , $w=0$ implies $e^0 = \Lambda x^0$ in (35). That Hamilton–Jacobi system then reads

$$(49) \quad \begin{aligned} \dot{x}^0 &= A_0 x^0, & x^0(s) &= \xi, \\ \dot{e}^0 &= (C_1' C_1 + \Lambda' B_2 B_2' \Lambda) x^0 - A_0' e^0, & e^0(t_1) &= 0. \end{aligned}$$

Consequently, with that definition there holds

$$(50) \quad \Lambda(s)\xi = e^0(s) = - \int_s^{t_1} \Phi_0(r, s)' (C_1' C_1 + \Lambda' B_2 B_2' \Lambda)(r) \Phi_0(r, s) \xi dr,$$

which, together with (7), shows that Λ is a negative-definite solution of (48).

Conversely, take any uniformly-bounded, negative-definite Ξ that satisfies (48) and such that the associated A_Ξ is exponentially stable. Substitute X for Λ and A_Ξ for A_0 in (49); then the L_2 solution of (49) defines an L_2 solution to (35) with $w=0$. The uniqueness of the solution to the latter implies that Ξ too is defined by (44). That is, our solution of (48) is unique. \square

⁷ Since a priori we do not know that (48) has a solution, nor that the solution is unique, neither that it is equal to Λ , as already defined above, we need to use an independent notation for the unknown in (48) (Ξ vis Λ) and for the evolution generated by the perturbed generator (Ψ vis Φ_0). Indeed, the observation that $\Xi = \Lambda$ and $\Psi = \Phi_0$ form the unique solution to (48) is the content of our proposition. The same policy applies to the notation used in (84).

We still owe the following. (Again, we note that a similar statement is established in [23].)

LEMMA 3.1.6. *If (1) is stabilizable then it is stabilizable by a zero-order state feedback.*

Proof. Our assumption is that there exists a dynamic compensator of the form (3) such that \mathcal{A} , as defined in (4), is an exponentially stable generator. We can rewrite (4) as

$$(51) \quad \mathcal{A} = \tilde{A} + \tilde{B}_2 \tilde{K}$$

(which is of the form of A_K) where

$$(52) \quad \tilde{A} \triangleq \begin{bmatrix} A & 0 \\ 0 & M \end{bmatrix}, \quad \tilde{B}_2 \triangleq \begin{bmatrix} B_2 & 0 \\ 0 & N \end{bmatrix}, \quad \tilde{K} \triangleq \begin{bmatrix} RC_2 & Q \\ C_2 & 0 \end{bmatrix}.$$

Denote also

$$(53) \quad \tilde{C}_1 \triangleq \begin{bmatrix} C_1 & 0 \\ 0 & I_P \\ 0 & 0 \end{bmatrix}, \quad \tilde{D}_{12} \triangleq \begin{bmatrix} D_{12} & 0 \\ 0 & 0 \\ 0 & I_Y \end{bmatrix}, \quad \tilde{x} \triangleq \begin{pmatrix} x \\ p \end{pmatrix}, \quad \tilde{u} \triangleq \begin{pmatrix} u \\ y \end{pmatrix}.$$

Substituting the $(\tilde{A}, \tilde{B}_2, \tilde{C}_1, \tilde{D}_{12})$ operators for their counterparts in the discussion above, we return to a setting of a system with a stabilizing state feedback. It involves a process in the product state space $X \times P$ with input from the product space $U \times Y$. Invoking Propositions 3.1.3–3.1.5, we thus conclude that the associated quadratic optimization problem $(\min_{\tilde{u}} \|\tilde{z} \triangleq \tilde{C}_1 \tilde{x} + \tilde{D}_{12} \tilde{u}\|_2)$ has a unique solution. Adding “~” to the various coefficients in the Hamilton–Jacobi system (35) and in the integral Riccati equation (48), we obtain a characterization of the optimal “control” in the augmented system, as described above. In particular, let us denote the solution of the augmented Riccati equation by $\tilde{\Lambda}$.

Now (save for the stabilizing feedback \tilde{K}) the augmented system coefficients are all block diagonal. In the case $w = 0$ we are thus actually dealing with two optimization problems, involving two noninteracting systems: The first involves a process in X , as described by the first and third equations in (1), while the other takes place in P . The solutions of the two problems are independent of each other: The optimal solution in X component is independent of the initial value of the P state, and vice versa. Using the equality (57), below (which as the reader will soon observe, must be satisfied in the augmented system), this implies that the $X \times P$ and $P \times X$ blocks of $\tilde{\Lambda}$ must vanish. That is $\tilde{\Lambda}$ is of the block diagonal form $\tilde{\Lambda} = \text{diag}(\Lambda, \Theta)$, where $\Lambda(t)$, $t \in (t_0, t_1)$ are negative-definite, uniformly bounded operators over X .

Moreover, as this implies that $\tilde{A}_0 \triangleq \tilde{A} - \tilde{B}_2 \tilde{B}_2' \tilde{\Lambda}$ is block diagonal, so is the evolution $\tilde{\Phi}_0$ that it generates. Substituting all those in the augmented version of (48) we find out that the X component of its solution, Λ , is a solution of (48) in its original form (i.e., with the “~” removed). Similarly, A_0 , which is the X component of \tilde{A}_0 , is stable. In particular, (1) is zero order stabilizable. \square

From now on we can take $K = B_2' \Lambda$ as our zero order stabilizing state feedback, and may substitute (37) by the control variable change

$$(54) \quad u = B_2' \Lambda x + v,$$

A_K by A_0 and Φ_K by Φ_0 , in the discussion above.

Suppose now that $w = 0$. Given any control input $u \in L_2$ that results in an L_2 state trajectory, let

$$(55) \quad u^\nabla \triangleq u - B_2' \Lambda x$$

be its momentary deviation from the optimal value (i.e., $u^\nabla = v$). The following is a standard, very useful observation.

PROPOSITION 3.1.7. *Given $t \geq s$, an initial state ξ , and an L_2 control u that result in a stable L_2 state trajectory,⁸ there holds*

$$(56) \quad \|\mathcal{Z}(s)(\xi, u, p)\|_{L_2[s,t]}^2 = \langle x(t), \Lambda(t)x(t) \rangle - \langle \xi, \Lambda(s)\xi \rangle + \|u^\nabla\|_{L_2[s,t]}^2.$$

In particular,

$$(57) \quad c^0(s; \xi, 0) = -\langle \xi, \Lambda(s)\xi \rangle.$$

Proof. In the finite dimensional case the simpler proof uses integration by parts. Differentiation is not allowed in the present context, whereby some toil is necessary.

We provide detail only for the case $t = t_1$. That is justified as follows: If $t_1 = \infty$ it is assumed that x is an L_2 function and that the limit $x(t_1) = 0$ exists. When $t_1 < \infty$ we recall that $\Pi_1(t_1) = 0$. Thus, in either case, the inner product term evaluated at $t = t_1$ vanishes from (56). The general form of (56), for any choice of t , is obtained by subtracting this equation's value over $[t, t_1]$ from its value over $[s, t_1]$.

We are interested in the quantities involved in

$$(58) \quad \begin{aligned} \|z\|^2 &= \|C_1 x\|^2 + \|u\|^2 \\ &= \|C_1 x\|^2 + \|u^\nabla\|^2 + 2\langle u^\nabla, B_2' \Lambda x \rangle + \|B_2' \Lambda x\|^2. \end{aligned}$$

Let us compute these quantities one by one. First, we have

$$(59) \quad \begin{aligned} &\|C_1 x\|_{L_2[s,t]}^2 \\ &= \int_s^t \left\langle C_1(r) \left(\Phi_0(r, s)\xi + \int_s^r \Phi_0(r, p)B_2(p)u^\nabla(p) dp \right), \right. \\ &\quad \left. C_1(r) \left(\Phi_0(r, s)\xi + \int_s^r \Phi_0(r, q)B_2(q)u^\nabla(q) dq \right) \right\rangle dr \\ &= \int_s^t \langle C_1(r)\Phi_0(r, s)\xi, C_1(r)\Phi_0(r, s)\xi \rangle dr \\ &\quad + 2 \int_s^t \left\langle C_1(r)\Phi_0(r, s)\xi, C_1(r) \int_s^r \Phi_0(r, q)B_2(q)u^\nabla(q) dq \right\rangle dr \\ &\quad + \int_s^t \left\langle C_1(r) \int_s^r \Phi_0(r, p)B_2(p)u^\nabla(p) dp, \right. \\ &\quad \left. C_1(r) \int_s^r \Phi_0(r, q)B_2(q)u^\nabla(q) dq \right\rangle dr \\ &= \left\langle \xi, \int_s^t \Phi_0'(r, s)C_1'(r)C_1(r)\Phi_0(r, s) dr \xi \right\rangle \\ &\quad + 2 \int_s^t \left\langle B_2'(q) \int_q^t \Phi_0'(r, q)C_1'(r)C_1(r)\Phi_0(r, q) dr \Phi_0(q, s)\xi, u^\nabla(q) \right\rangle dq \\ &\quad + 2 \int_s^t \left\langle B_2'(q) \int_q^t \Phi_0'(r, q)C_1'(r)C_1(r)\Phi_0(r, q) dr \right. \\ &\quad \left. \cdot \int_s^q \Phi_0(q, p)B_2(p)u^\nabla(p) dp, u^\nabla(q) \right\rangle dq \end{aligned}$$

⁸ That is, if $t_1 = +\infty$ then $\exists \lim_{t \rightarrow +\infty} x(t) = 0$.

$$= \left\langle \xi, \int_s^t \Phi'_0(r, s) C'_1(r) C_1(r) \Phi_0(r, s) dr \xi \right\rangle \\ + 2 \int_s^t \left\langle B'_2(q) \int_q^t \Phi'_0(r, q) C'_1(r) C_1(r) \Phi_0(r, q) dr x(q), u^\nabla(q) \right\rangle dq.$$

Similarly,

$$\|B'_2 \Lambda x\|^2 = \left\langle \xi, \int_s^{t_1} \Phi'_0(r, s) (\Lambda B_2 B'_2 \Lambda)(r) \Phi_0(r, s) dr \xi \right\rangle \\ (60) \quad + 2 \int_s^{t_1} \left\langle B'_2(q) \int_q^{t_1} \Phi'_0(r, q) (\Lambda B_2 B'_2 \Lambda)(r) \Phi_0(r, q) dr x(q), u^\nabla(q) \right\rangle dq.$$

Adding (59) and (60) we get

$$(61) \quad \|C_1 x\|^2 + \|B'_2 \Lambda x\|^2 = -\langle \xi, \Lambda(s) \xi \rangle - 2 \int_s^{t_1} \langle B'_2(q) \Lambda(q) x(q), u^\nabla(q) \rangle dq.$$

Substituting (61) into (58) we get (56) (without the inner product term in $t = t_1$, which vanishes as explained above). \square

3.2. The min in (27). Throughout this section we assume that $\gamma > \gamma_0$ (where γ_0 is equal to the optimal value in (1)). We denote

$$(62) \quad J^0(s; \xi, w) \triangleq J(s; \xi, \mathcal{U}^0(s)(\xi, w), w) \\ = \gamma^2 \|w\|^2 - \|\mathcal{L}^0(s)(\xi, w)\|^2$$

and search for w^* such that

$$(63) \quad J^0(s; \xi, w^*) = J^*(s; \xi) \triangleq \inf_w J^0(s; \xi, w).$$

The solution of this problem will follow, proposition by proposition, the analysis from § 3.1, with J^0 substituting for c as the quadratic cost, and where w substitutes u as the optimization variable. Some modifications will be needed, of course, due to the fact that c were positive definite whereas J^0 is not. These modifications are possible exactly when γ is strictly suboptimal, as assumed here.

PROPOSITION 3.2.1. Assume that $\gamma > \gamma_0$. Then (i) for all s and w

$$(64) \quad J^0(s; 0, w) \geq \delta^2 \|w\|^2$$

(where δ is as in (25)). (ii) There exists $\mu > 0$ such that

$$(65) \quad J^0(s; \xi, w) \geq -\mu \|\xi\|^2$$

for all s , w , and ξ . In particular,

$$(66) \quad J^*(s; 0) = 0$$

and

$$(67) \quad 0 > J^*(s; 0) > -\mu \|\xi\|^2$$

for $\xi \neq 0$.

Proof. (i) By Observation 3.0.1. there exists an internally stabilizing compensator \mathcal{K} such that (25) holds:

$$(68) \quad \gamma^2 \|w\|^2 - \|\mathcal{T}_{\mathcal{K}}(s)w\|^2 \geq \delta^2 \|w\|^2.$$

By definition,

$$(69) \quad J^0(s; 0, w) \geq \gamma^2 \|w\|^2 - \|\mathcal{T}_{\mathcal{K}}(s)w\|^2.$$

Hence this part of the proposition.

(ii) There holds

$$\begin{aligned}
 J^0(s; \xi, w) &= \gamma^2 \|w\|^2 - \|\mathcal{Z}^0(s)(\xi, w)\|^2 \\
 &= \gamma^2 \|w\|^2 - \|\mathcal{Z}^0(s)(\xi, 0) + \mathcal{Z}^0(0, w)\|^2 \\
 &\geq \gamma^2 \|w\|^2 - (\|\mathcal{Z}^0(s)(\xi, 0)\| + \|\mathcal{Z}^0(0, w)\|)^2 \\
 &= \gamma^2 \|w\|^2 - \|\mathcal{Z}^0(s)(0, w)\|^2 \\
 &\quad - 2\|\mathcal{Z}^0(s)(\xi, 0)\| \|\mathcal{Z}^0(0, w)\| - \|\mathcal{Z}^0(s)(\xi, 0)\|^2 \\
 &\geq \delta^2 \|w\|^2 - 2\|\mathcal{Z}^0(s)(\xi, 0)\| \|\mathcal{Z}^0(0, w)\| - \|\mathcal{Z}^0(s)(\xi, 0)\|^2 \\
 &\geq \delta^2 \|w\|^2 - \|\mathcal{Z}^0(s)\|^2 (2\|\xi\| \|w\| + \|\xi\|^2) \\
 &\geq \delta^2 \|w\|^2 - \theta^2 (2\|\xi\| \|w\| + \|\xi\|^2),
 \end{aligned}
 \tag{70}$$

where θ is as in Proposition 3.1.1. The right-most term in (70) is a quadratic form in $\|w\|$ which minimum is $-\theta^2(\theta^2/\delta^2 + 1)\|\xi\|^2$. So $\mu = \theta^2(\theta^2/\delta^2 + 1)$ satisfies the claim.

Substituting $w=0$ in (62) we see that $\inf_w J_0(s; \xi, w)$ is nonpositive; in fact, due to (7), the infimum is negative unless $\xi=0$. This completes the proof. \square

PROPOSITION 3.2.2. *Assume that $\gamma > \gamma_0$. Then, given ξ , there exists a unique disturbance $w^* = \mathcal{W}^*(s)(\xi)$ that attains the minimum in (63). Moreover, the operators $\mathcal{W}^*(s)$ are uniformly bounded for all choices of s .*

Proof. By (67) the infimum is finite. Moreover, it follows from that inequality together with (70) that disturbances that make J^0 approach its infimum are uniformly norm bounded; indeed, (70) coupled with the requirement $J^0 \leq 0$ implies that w satisfies

$$\|w\| \leq \frac{\theta(\theta + \sqrt{\theta^2 + \delta^2})}{\delta^2} \|\xi\|.
 \tag{71}$$

By (64) there holds

$$\begin{aligned}
 \|w_1 - w_2\|^2 &\leq \frac{1}{\delta^2} J^0(s; 0, w_1 - w_2) \\
 &= \frac{2}{\delta^2} \left(J^0(s; \xi, w_1) + J^0(s; \xi, w_2) - 2J^0\left(s; \xi, \frac{1}{2}(w_1 + w_2)\right) \right) \\
 &\leq \frac{2}{\delta^2} (J^0(s; \xi, w_1) + J^0(s; \xi, w_2) - 2J^*(s; \xi)).
 \end{aligned}
 \tag{72}$$

As in the proof of Proposition 3.1.2, (71) and (72) imply existence and uniqueness of w^* . Equation (71) also assures uniform boundedness of $\mathcal{W}^*(s)$. \square

We denote $\mathcal{U}^*(s)(\xi) \triangleq \mathcal{U}^0(s)(\xi, \mathcal{W}^*(s)(\xi))$, $\mathcal{X}^*(s)(\xi) \triangleq \mathcal{X}^0(s)(\xi, \mathcal{W}^*(s)(\xi))$, $\mathcal{E}^*(s)(\xi) \triangleq \mathcal{E}^0(s)(\xi, \mathcal{W}^*(s)(\xi))$ and $\mathcal{Z}^*(s)(\xi) \triangleq \mathcal{Z}^0(s)(\xi, \mathcal{W}^*(s)(\xi))$. It immediately follows that all these mappings define uniformly bounded operators.

PROPOSITION 3.2.3. *Assume that $\gamma > \gamma_0$. Then there exists a unique L_2 solution to the following Hamilton-Jacobi boundary-value problem:⁹*

$$\begin{aligned}
 \dot{x} &= Ax + \left(B_2 B_2' - \frac{1}{\gamma^2} B_1 B_1' \right) e, & x(s) &= \xi, \\
 \dot{e} &= C_1' C_1 x - A' e, & e(t_1) &= 0.
 \end{aligned}
 \tag{73}$$

That solution is given by $x^ = \mathcal{X}^*(s)(\xi)$, $e^* = \mathcal{E}^*(s)(\xi)$. In particular, the latter are bounded linear operators.*

⁹ Recall that when $t_1 = +\infty$ the terminal condition reads $\exists \lim_{t \rightarrow +\infty} e(t) = 0$.

Proof. As in the proof of Proposition 3.1.3, we first convert (1) to a stable system by use of a zero-order compensator. That is done by introducing the control change of variables

$$(74) \quad u = B'_2 \Lambda x + v.$$

In terms of (74), the first and last equations in (1) read

$$(75) \quad \dot{x} = A_0 x + B_1 w + B_2 v, \quad z = (C_1 + D_{12} B'_2 \Lambda) x + D_{12} v.$$

We denote by $v^0 = \mathcal{V}^0(s)(\xi, w)$ the value of v associated via (74) with the optimal control, and by abuse of notation, with $z = \mathcal{Z}(s)(\xi, v, w)$ and $z^0 = \mathcal{Z}^0(s)(\xi, w)$ the output and optimal output mappings in (75).

Given any two L_2 disturbances, w^+ and w , there holds

$$(76) \quad \begin{aligned} J^0(s; \xi, w^+ + w) &= J^0(s; \xi, w^+) + 2(\gamma^2 \langle w^+, w \rangle - \langle \mathcal{Z}^0(s)(\xi, w^+), \mathcal{Z}^0(s)(0, w) \rangle) \\ &\quad + J^0(s; 0, w) \end{aligned}$$

Since the last term on the right-hand side is nonnegative (by (64)), w^+ is minimizing if and only if the inner product term vanishes for all choices of w .

Expanding on that term we have

$$(77) \quad \begin{aligned} &\gamma^2 \langle w^+, w \rangle - \langle \mathcal{Z}^0(s)(\xi, w^+), \mathcal{Z}^0(s)(0, w) \rangle \\ &= \gamma^2 \langle w^+, w \rangle - \langle \mathcal{Z}^0(s)(\xi, w^+), \mathcal{Z}(s)(0, \mathcal{V}^0(s)(0, w), w) \rangle \\ &= \gamma^2 \langle w^+, w \rangle - \langle \mathcal{Z}^0(s)(\xi, w^+), \mathcal{Z}(s)(0, 0, w) \rangle \\ &\quad - \langle \mathcal{Z}^0(s)(\xi, w^+), \mathcal{Z}(s)(0, \mathcal{V}^0(s)(0, w), 0) \rangle. \end{aligned}$$

As shown in the proof of Proposition 3.1.3, the last term on the right-hand side of (77) vanishes. (Indeed, since $\mathcal{Z}^0(s)(\xi, w^+)$ is the output associated with the optimal value v^0 of v for the disturbance choice $w = w^+$, we concluded from (39) that (40) holds; in the present notation this means that

$$(78) \quad \langle \mathcal{Z}^0(s)(\xi, w^+), \mathcal{Z}(s)(0, v, 0) \rangle = 0$$

for all choices of v .)

Straightforward computation shows that the rest of the right-hand side of (77) vanishes for all choices of $w \in L_2$ if and only if

$$(79) \quad w^+ = -\frac{1}{\gamma^2} B'_1 e^+,$$

where

$$(80) \quad e^+(t) = - \int_t^{t_1} \Phi'_0(r, t) (C'_1 C_1 x^+ + \Lambda B_2 u^+)(r) dr$$

and

$$(81) \quad x^+ = \mathcal{X}^0(s)(\xi, w^+), \quad u^+ = \mathcal{U}^0(s)(\xi, w^+).$$

Substituting (74) for (37) as the definition of the zero-order stabilizing feedback, (80) is equal to the explicit expression in (42) for \mathcal{E}^0 ; namely, it reads $e^+ = \mathcal{E}^0(s)(\xi, w^+)$. This establishes existence of a solution to (73).

Moreover, we have shown that any solution of (73) defines the unique solution to the min-max problem (27) via (41) and (79). Thus, even if the solution to (73) were not unique, the functions x , $B_1'e$, and $B_2'e$ it creates are uniquely determined. It thus remains to be shown that when all the latter vanish then $e=0$ as well, which we established in the proof of Proposition 3.1.3. \square

Denote

$$(82) \quad \Pi_1(t)\xi \triangleq (\mathcal{E}^*(t)(\xi))(t)$$

(following from (80), these too are uniformly bounded linear operators) and let

$$(83) \quad A_1 = A + \left(B_2 B_2' - \frac{1}{\gamma^2} B_1 B_1' \right) \pi_1$$

as defined in Theorem I.

PROPOSITION 3.2.4. A_1 is exponentially stable.

The proof is completely analogous to the proof of Proposition 3.1.4, and is thus left out.

PROPOSITION 3.2.5. $\Xi = \Pi_1$ is the unique L_∞ bounded-operator-valued, negative-definite solution of the integral Riccati equation (8)

$$(84) \quad \Xi(s) = - \int_s^{t_1} \Psi(t, s)' \left[C_1' C_1 + \Xi \left(B_2 B_2' - \frac{1}{\gamma^2} B_1 B_1' \right) \Xi \right] (t) \Psi(t, s) ds,$$

where Ψ is the exponentially stable evolution generated by $A_\Xi \triangleq A + (B_2 B_2' - \frac{1}{\gamma^2} B_1 B_1') \Xi$.

Proof. The proof that Π_1 is the unique selfadjoint solution to our Riccati equation is essentially the same as its counterpart in the proof Proposition 3.1.5. Since the integrand in (84) contains both positive and negative parts, it remains to be established that Π_1 is negative definite.

Indeed, by Proposition 3.2.3 and the definition (82) there holds $e^* = \Pi_1 x^*$, whereby the optimal min-max inputs satisfy $u^* = B_2' \Pi_1 x^*$ and $w^* = -(1/\gamma^2) B_1' \Pi_1 x^*$. Substituting the entire right-hand side of (84) for $\Pi_1(s)$, the following equality is then obtained:

$$(85) \quad J^*(s; \xi) = \langle \xi, \Pi_1(s) \xi \rangle.$$

By Proposition 3.2.1, the left-hand side of (85) is a negative expression in ξ , which completes the proof. \square

We denote by

$$(86) \quad u^\Delta \triangleq u - B_2' \Pi_1 x, \quad w^\Delta \triangleq w + \frac{1}{\gamma^2} B_1' \Pi_1 x$$

the momentary deviations of the control and disturbance in (1) from their min-max optimal values. The following is a direct counterpart of Proposition 3.1.7, where the guideline of the proof can be found.

PROPOSITION 3.2.6. Given an initial time s and an initial state ξ and $t \geq s$, if u , w , and x in (1) are all L_2 functions then

$$(87) \quad \gamma^2 \|w\|_{L_2[s,t]}^2 - \|z\|_{L_2[s,t]}^2 = \langle \xi, \Pi_1(s) \xi \rangle - \langle x(t), \Pi_1(t) x(t) \rangle + \gamma^2 \|w^\Delta\|_{L_2[s,t]}^2 - \|u^\Delta\|_{L_2[s,t]}^2.$$

3.3. Completion of the proof of (c).

PROPOSITION 3.3.1. Assume that $\gamma > \gamma_0$. Then the zero-order state feedback $u = B_2' \Pi_1 x$ is internally stabilizing, and it guarantees strict γ -suboptimality; namely, $\|\mathcal{T}_x(t_0)\| < \gamma$. Conversely, if there exists a solution to the Riccati equation (8) with the properties described in part (a) of Theorem I, then $\gamma > \gamma_0$.

Proof. Assume that $\gamma > \gamma_0$. Then there exists Π_1 , satisfying the requirement in part (a) of Theorem I; in particular, such that A_1 is stable. The remainder of the proof depends only on the existence of such Π_1 . Set

$$(88) \quad A_3 \triangleq A + B_2 B_1' \Pi_1$$

and let Φ_3 be the evolution generated by A_3 . In the case $t_1 = +\infty$ we must establish exponential stability of Φ_3 . Indeed, plugging $w = 0$ and our feedback control into (87), and invoking assumption (7) (via (30)) we get

$$(89) \quad \begin{aligned} \varepsilon \|x\|_{L_2[s, t]}^2 &\leq \|z\|_{L_2[s, t]}^2 \\ &= -\langle \xi, \Pi_1(s)\xi \rangle + \langle x(t), \Pi_1(t)x(t) \rangle - \gamma^2 \|w^\Delta\|_{L_2[s, t]}^2 \\ &\leq -\langle \xi, \Pi_1(s)\xi \rangle \leq \|\Pi_1\|_\infty \|\xi\|^2 \end{aligned}$$

for all $t \geq s$ in the closed-loop system. Therefore

$$(90) \quad \forall \xi, \quad \int_s^{t_1} \|\Phi_3(r, s)\xi\|^2 dr \leq \frac{\|\Pi_1\|_\infty}{\varepsilon} \|\xi\|^2$$

for all s . Stability follows now from Lemma 2.2.

Allowing $w \neq 0$ with $\xi = 0$ and $t = t_1$, our state feedback turns (87) into

$$(91) \quad \gamma^2 \|w\|^2 - \|z\|^2 = \gamma^2 \|w^\Delta\|^2.$$

The closed-loop mapping $w^\Delta \rightarrow w$ is governed by a stable system

$$(92) \quad \dot{x} = A_1 x + B_1 w^\Delta, \quad w = w^\Delta - \frac{1}{\gamma^2} B_1 B_1' x.$$

In particular, that mapping is bounded. There exists thereby some $\delta \neq 0$ such that

$$(93) \quad \gamma^2 \|w^\Delta\|^2 \geq \delta^2 \|w\|^2.$$

Couple (91) with (93) to get

$$(94) \quad \gamma^2 \|w\|^2 - \|z\|^2 \geq \delta^2 \|w\|^2.$$

By Observation 3.0.1 the desired inequality $\|\mathcal{T}_{\mathcal{K}}(t_0)\| < \gamma$ follows. \square

3.4. Conjugation. Let a stabilizing feedback $u = \mathcal{K}y$ be given. Taking $t = t_1$ and $\xi = 0$, equation (87) suggests that the closed-loop mapping $\mathcal{T}_{\mathcal{K}}: w \rightarrow z$ has norm smaller than γ if and only if the mapping $\mathcal{T}_{\mathcal{K}}^\Delta: w^\Delta \rightarrow u^\Delta$ is so bounded. Keeping a cap on the latter is an estimation problem: we want the closed-loop observation-based control to be as good an estimate as possible of the state feedback $u = B_2' \Pi_1 x$.

The following system¹⁰ governs the mapping $\mathcal{T}_{\mathcal{K}}^\Delta: w^\Delta \rightarrow u^\Delta$:

$$(95) \quad \begin{aligned} \dot{x} &= \left(A - \frac{1}{\gamma^2} B_1 B_1' \Pi_1 \right) x + B_1 w^\Delta + B_2 u, \\ u^\Delta &= -B_2' \Pi_1 x + u, \\ y &= C_2 x + D_{21} w^\Delta, \quad u = \mathcal{K}y. \end{aligned}$$

It does not adhere with the first line of assumption (7), but its conjugate does. We thus obtain a second Riccati equation as a necessary condition, in terms of the conjugate

¹⁰ Note that by assumption (7) there holds $D_{21} w = D_{21} w^\Delta$, an equality we use in the observation term, in (94).

system. That will follow the tradition of converting optimal estimation problems into optimal input problems in a conjugate setting.

As a first and crucial step, we need the following proposition.

PROPOSITION 3.4.1. *Let Π_1 , be as stated in Theorem I, and let u^Δ and w^Δ be defined by (86) in terms of that Π_1 . Then the observation feedback $u = \mathcal{K}y$ is internally stabilizing and strictly γ -suboptimal in (1) if and only if it is so in (95).*

Proof. As a first step assume that the observation-feedback $u = \mathcal{K}y$ is internally stabilizing in both (1) and (95). It is easy to show that $u = \mathcal{K}y$ is then strictly γ -suboptimal in (1) if and only if it is so in (95). Indeed, under the stability assumption both the closed-loop mappings $w \leftrightarrow w^\Delta$ are continuous. Given the zero initial state we then have

$$(96) \quad \alpha \|w\|^2 \leq \|w^\Delta\|^2 \leq \beta \|w\|^2$$

for some positive α and β . Consider now (87) with $s = t_0$, $t = t_1$, and $\xi = 0$. Invoking Observation 3.0.1, $u = \mathcal{K}y$ is strictly γ -suboptimal in (1) if and only if the left-hand side of (87) is bounded below by $\delta^2 \|w\|^2$, which, by (96), is true if and only if the right-hand side of (87) is bounded below by $\eta^2 \|w^\Delta\|^2$ (for some appropriate positive δ and η), and, in turn, if and only if $u = \mathcal{K}y$ is strictly γ -suboptimal in (95).

For the main part of the proof, assume that the observation-feedback $u = \mathcal{K}y$ is internally stabilizing and strictly γ -suboptimal in (1). We then show that $u = \mathcal{K}y$ is stabilizing in (95) as well. (The proof of the converse proposition, going from stability and γ -suboptimality in (95) to stability in (1) is completely similar.) The main idea in the proof is the following. If the disturbance is such that $w^\Delta = 0$ then, by (87), the term $\gamma^2 \|w\|_{L_2[s, t_1]}^2 - \|z\|_{L_2[s, t_1]}^2$ is nonpositive in the closed-loop system. Under the assumption that the claim fails and \mathcal{K} is not internally stabilizing in (95) we shall find such a disturbance that makes the term positive; a contradiction. It will require some work.

Let (3) be a realization of \mathcal{K}

$$(97) \quad \dot{p} = Mp + Ny, \quad u = Qp + Ry,$$

let the exponentially stable closed-loop generator be as in (4)

$$(98) \quad \mathcal{A} = \begin{bmatrix} A + B_2 RC_2 & B_2 Q \\ NC_2 & M \end{bmatrix},$$

and let Φ be the evolution generated by \mathcal{A} over $\mathbf{X} \times \mathbf{P}$. We denote also

$$\mathcal{B} \triangleq \begin{bmatrix} B_1 + B_2 RD_{21} \\ ND_{21} \end{bmatrix}, \quad \mathcal{A}_0 \triangleq \mathcal{A} + \mathcal{B} \begin{bmatrix} -\frac{1}{\gamma^2} B_1' \Pi_1 & 0 \end{bmatrix}.$$

The claim is that the evolution Φ_0 , generated by \mathcal{A}_0 , is exponentially stable.

The equation

$$(99) \quad \dot{\tilde{x}} = \mathcal{A}_0 \tilde{x}$$

is equivalent to

$$(100) \quad \dot{\tilde{x}} = \mathcal{A} \tilde{x} + \mathcal{B} w$$

subject to the feedback

$$(101) \quad w = -\frac{1}{\gamma^2} B_1' \Pi_1 x,$$

where $\tilde{x} = [\tilde{x}_p]$. By Lemma 2.2, if the current proposition is false then the ratios

$$(102) \quad \frac{\|\tilde{x}\|_{L_2[s,t]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}}, \quad \tilde{x}(s) \in \mathbf{X} \times \mathbf{P}, \quad t_0 < s < t < t_1,$$

as governed by (99), are not uniformly bounded.

Suppose therefore that our proposition is false, and that (99) is not an exponentially stable equation. If the ratios

$$(103) \quad \frac{\|x\|_{L_2[s,t]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}}, \quad \tilde{x}(s) \in \mathbf{X} \times \mathbf{P}, \quad t_0 < s < t < t_1$$

were uniformly bounded, then by (101) so would be the ratios

$$(104) \quad \frac{\|w\|_{L_2[s,t]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}}, \quad \tilde{x}(s) \in \mathbf{X} \times \mathbf{P}, \quad t_0 < s < t < t_1.$$

Consequently, the stability of \mathcal{A} in (100) will impose a uniform bound on (102), in contrast to our assumption. Thus, if \mathcal{H} is not internally stabilizing in (95) then (103) is unbounded. We continue on that premise.

Let now $t_0 < s < t < t_1$ and $\tilde{x}(s)$ be given; then let \tilde{x} be the associated trajectory of (99) and denote $w \triangleq -(1/\gamma^2)B_1^*\Pi_1x$, $u \triangleq [RC_2 \quad Q]\tilde{x}$, and $z \triangleq C_1x + D_{12}u$. All these functions are thus far defined along $[s, t]$. Along the following interval $[t, t_1]$ we consider the optimization problem

$$(105) \quad \min_{w \in L_2} (\gamma^2 \|w\|^2 - \|z\|^2),$$

subject to (99) and the initial state $\tilde{x}(t)$, as already determined. Here again, the control and output are defined in terms of the realization of $\mathcal{H}: u \triangleq [RC_2 \quad Q]\tilde{x} + RD_{21}w$ and $z \triangleq C_1x + D_{12}u$. By arguments that are completely similar to those used in §§ 3.1 and 3.2 it follows that a unique solution to (105) exists, which we denote simply by w .

Now w is defined along the entire interval $[s, t_1]$, associated with a state trajectory \tilde{x} , a control u , and an output z . The observation that the ratios (102), (103), and (104) are unbounded implies that so are the ratios

$$(106) \quad \frac{\|\tilde{x}\|_{L_2[s,t_1]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}}, \quad \frac{\|x\|_{L_2[s,t_1]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}}, \quad \frac{\|w\|_{L_2[s,t_1]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}}, \quad \frac{\|z\|_{L_2[s,t_1]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}},$$

$$\tilde{x}(s) \in \mathbf{X} \times \mathbf{P}, \quad t_0 < s < t_1,$$

as derived by the procedure above.

We are going to construct yet two more sets of trajectories. Given the initial state $\tilde{x}(s)$, let \tilde{x}_σ be the following free motion of (100) (i.e., $\tilde{x}_\sigma(t) = \Phi(t, s)\tilde{x}(s)$), set $u_\sigma \triangleq [RC_2 \quad Q]\tilde{x}_\sigma$ and $z_\sigma \triangleq C_1x_\sigma + D_{12}u_\sigma$. Then set $\tilde{x}_\nu \triangleq \tilde{x} - \tilde{x}_\sigma$, $u_\nu \triangleq u - u_\sigma$, and $z_\nu \triangleq z - z_\sigma$.

Since Φ is exponentially stable, the ratios

$$(107) \quad \frac{\|z_\sigma\|_{L_2[s,t_1]}}{\|\tilde{x}(s)\|_{\mathbf{X} \times \mathbf{P}}}, \quad \tilde{x}(s) \in \mathbf{X} \times \mathbf{P}, \quad t_0 < s < t_1,$$

are uniformly bounded. It is therefore possible to select our parameters so that z_ν be increasingly the dominant part of z . That is, we can select s , t , and $\tilde{x}(s)$ so that

$$(108) \quad (1 - \lambda^2)\|z_\nu\|^2 \leq \|z\|^2 \leq (1 + \lambda^2)\|z_\nu\|^2$$

with arbitrarily small λ .

Let us carefully examine the trajectory x_ν . As follows from the definition above, we have

$$\begin{bmatrix} x_\nu(s) \\ p_\nu(s) \end{bmatrix} = \tilde{x}_\nu(s) = 0.$$

Thus, u_ν is just the feedback control $u_\nu = \mathcal{K}y_\nu$ ($\triangleq \mathcal{K}(C_2x_\nu + D_{21}w)$) and $z_\nu = \mathcal{Z}_{\mathcal{K}}(s)(0, u_\nu, w)$ (equal to the output associated with the disturbance w , the zero initial state, and the closed-loop control law $u = \mathcal{K}y$). Since we assume that \mathcal{K} is strictly γ -suboptimal, Observation 3.0.1. tells us that there holds

$$(109) \quad \gamma^2 \|w\|^2 - \|z_\nu\|^2 \geq \delta^2 \|w\|^2$$

with some nonzero δ that can be selected independent of our present constructions.

Combining (108) and (109) we get

$$(110) \quad \gamma^2 \|w\|^2 - \|z\|^2 \geq \gamma^2 \|w\|^2 - (1 + \lambda^2) \|z_\nu\|^2 \geq (\delta^2 - \lambda^2 \gamma^2) \|w\|^2.$$

As γ and δ are fixed and λ can be chosen arbitrarily small, we can select our parameters so that the right-hand side of (110) be positive; indeed, we can make it arbitrarily large.

Our next move is to show that a positive left-hand side in (110) is impossible. For that purpose we define $x^* \triangleq x$, $u^* \triangleq u$, and $z^* \triangleq z$ along $[s, t]$. Then we extend these trajectories via $x^*(\cdot) \triangleq \Phi_1(\cdot, t)x^*(t)$, $u^* \triangleq B_2' \Pi_1 x^*$, and $w^* \triangleq -(1/\gamma^2) B_1' \Pi_1 x^*$ along $[t, t_1]$. Since Φ_1 is exponentially stable, all these are L_2 trajectories. Moreover, the associated $w^{*\Delta}$ vanishes throughout. In particular, by (87) there holds

$$(111) \quad \gamma^2 \|w^*\|_{L_2[s, t_1]}^2 - \|z^*\|_{L_2[s, t_1]}^2 \leq 0.$$

As established in §§ 3.1 and 3.2, along $[t, t_1]$ we have

$$\begin{aligned} & \gamma^2 \|w^*\|_{L_2[t, t_1]}^2 - \|z^*\|_{L_2[t, t_1]}^2 \\ &= \min_{f \in L_2} \max_{g \in L_2} (\gamma^2 \|f\|_{L_2[t, t_1]}^2 - \|\mathcal{Z}(t)(x(t), g, f)\|_{L_2[t, t_1]}^2) \\ (112) \quad & \geq \min_{f \in L_2} (\gamma^2 \|f\|_{L_2[t, t_1]}^2 - \|\mathcal{Z}(t)(x(t), g = [RC_1 \quad Q]\tilde{x} + RD_{21}f, f)\|_{L_2[t, t_1]}^2) \\ &= \gamma^2 \|w\|_{L_2[t, t_1]}^2 - \|z\|_{L_2[t, t_1]}^2. \end{aligned}$$

We recall that, by definition, $w^* = w$ and $z^* = z$ along $[s, t]$. Thus the inequality between the right-most and the left-most terms in (112) can be extended to the entire interval $[s, t_1]$. Joining this inequality with (110) and (111), the contradiction

$$\begin{aligned} & 0 \geq \gamma^2 \|w^*\|_{L_2[s, t_1]}^2 - \|z^*\|_{L_2[s, t_1]}^2 \\ (113) \quad & \geq \gamma^2 \|w\|_{L_2[s, t_1]}^2 - \|z\|_{L_2[s, t_1]}^2 \\ & > (\delta^2 - \lambda^2 \gamma^2) \|w\|_{L_2[s, t_1]}^2 > 0 \end{aligned}$$

follows. This completes the proof. \square

Now we are almost ready to conjugate (95). As a matter of convenience (and personal preference)¹¹ we introduce the following notion of conjugation. Let Ω be the time reversal operator, applied to functions over \Re : $(\Omega f)(t) \triangleq f(-t)$. Given an operator $\mathcal{E}: L_2(t_0, t_1) \rightarrow L_2(t_0, t_1)$ we define its *conjugate* by $\mathcal{E}^\# = \Omega \mathcal{E}' \Omega: L_2(-t_1, -t_0) \rightarrow L_2(-t_1, -t_0)$ (where \mathcal{E}' is the usual adjoint). Obviously $\mathcal{E}^{\#\#} = \mathcal{E}$ and $\|\mathcal{E}\| = \|\mathcal{E}^\#\|$, and

¹¹ This way the conjugate system will be causal, instead of anticausal, and we will be able to directly draw on the results of the previous sections.

if \mathcal{E} is a multiplication operator (i.e., $(\mathcal{E}f)(t) = E(t)f(t)$, $t \in (t_0, t_1)$, for some operator valued function $E(t): \mathbb{F} \rightarrow \mathbb{F}$) then $(\mathcal{E}^\# g)(t) = E'(-t)g(t)$, $t \in (-t_1, -t_0)$. Finally, if E generates an exponentially stable evolution, $\Psi(t, s)$, $t_1 > t \geq s > t_0$, then $E^\#$ generates the exponentially stable evolution $\Psi'(-s, -t)$, $-t_0 > t \geq s > -t_1$.

Let $\mathcal{T}_{\mathcal{K}}^{\Delta\#}$ be the input-output operator defined via the realization (95), as above. Then its conjugate, $\mathcal{T}_{\mathcal{K}^\#}^{\Delta\#}$, is realized by

$$\begin{aligned} \dot{\hat{x}} &= \left(A - \frac{1}{\gamma^2} B_1 B_1' \Pi_1 \right)^\# \hat{x} - \Pi_1^\# B_2^\# \hat{w} + C_2^\# \hat{u}, \\ \hat{z} &= B_1^\# \hat{x} + D_{21}^\# \hat{u}, \\ \hat{y} &= B_2^\# \hat{x} + \hat{w}, \quad \hat{u} = \mathcal{K}^\# \hat{y} \end{aligned} \quad (114)$$

where $\mathcal{K}^\#$ has the realization

$$\dot{\hat{p}} = M^\# \hat{p} + Q^\# \hat{y}, \quad \hat{u} = N^\# \hat{p} + R^\# \hat{y}. \quad (115)$$

The following is an immediate result of Proposition 3.4.1.

COROLLARY 3.4.2. *The feedback $u = \mathcal{K}y$ is internally stabilizing and strictly γ -suboptimal in (1) if and only if $\hat{u} = \mathcal{K}^\# \hat{y}$ as in (114).*

Applying the analysis of §§ 3.1 and 3.2 to (114) we thus get the following corollary.

COROLLARY 3.4.3. *If $\gamma > \gamma_0$ in (1) and Π_1 is as described above then there exists also Π_2 , a solution to the Riccati equation (9), as stated in Theorem I.*

Proof. All it takes is to note that, due to the assumptions in the second line of (7), the open-loop part of (114) satisfies assumptions in the first line of (7), which we used in § 3.1, and that the role of the conjugated equation (9) in (114) is exactly that of (8) in (1). \square

This completes the proof of necessity in part (a) of Theorem I.

3.5. The parametrization of γ -suboptimal compensators.

PROPOSITION 3.5.1. *Every internally stabilizing, strictly γ -suboptimal compensator admits a realization of the form (10):*

$$\begin{aligned} \dot{p} &= (A_1 + \Pi_2 C_2' C_2) p + \Pi_2 C_2' y - \left(I + \frac{1}{\gamma^2} \Pi_2 \Pi_1 \right) B_2 v, \\ u &= -B_2' \Pi_1 p + v, \\ q &= C_2 p + y, \quad v = \mathcal{K}_0 q, \end{aligned} \quad (116)$$

where \mathcal{K}_0 has an exponentially stable realization and $\|\mathcal{K}_0\| < \gamma$.

Proof. Suppose that \mathcal{K} is an internally stabilizing, strictly γ -suboptimal compensator in (1). We define \hat{u}^Δ and \hat{w}^Δ in (114) following the same logic used in the analysis of (1). Here the appropriate definitions are

$$\hat{u}^\Delta = \hat{u} - C_2^\# \Pi_2^\# \hat{x}, \quad \hat{w}^\Delta = \hat{w} - \frac{1}{\gamma^2} B_2^\# \Pi_1^\# \Pi_2^\# \hat{x}. \quad (117)$$

Let then $\mathcal{K}_0^\#$ stand for the closed-loop mapping $\hat{w}^\Delta \rightarrow \hat{u}^\Delta: L_2(-t_1, -t_0) \rightarrow L_2(-t_1, -t_0)$. Following from Proposition 3.4.1 (applied to (114)), this mapping admits an exponentially stable realization and its norm is strictly smaller than γ .

Consider the following compensator $\mathcal{K}_1^\#$ as an alternative to $\mathcal{K}^\#$ in (114):

$$(118) \quad \begin{aligned} \dot{\hat{p}} &= (A_1 + \Pi_2 C_2' C_2)^\# \hat{p} - \Pi_1^\# B_2^\# \hat{y} + C_2^\# \hat{v}, \\ \hat{u} &= C_2^\# \Pi_2^\# \hat{p} + \hat{v}, \\ \hat{q} &= -B_2^\# \left(I + \frac{1}{\gamma^2} \Pi_2 \Pi_1 \right)^\# \hat{p} + \hat{y}, \quad \hat{v} = \mathcal{H}_0^\# \hat{q}. \end{aligned}$$

Setting $\hat{e} \triangleq \hat{x} - \hat{p}$ it easily follows that \hat{e} satisfies a homogeneous ordinary differential equation (ODE). In particular, if both the initial states in (114) and in (118) are zero then $\hat{x} = \hat{p}$. Consequently, then¹² $\hat{q} = \hat{w}^\Delta$ and finally $\hat{u} = C_2^\# \Pi_2^\# \hat{x} + \mathcal{H}_0^\# \hat{w}^\Delta$. Following from the definition of \mathcal{H}_0 this means that then $\hat{u} = \mathcal{K}^\# \hat{y}$.

In short, $\mathcal{K}^\#$ and $\mathcal{K}_1^\#$ coincide over the range of the closed-loop mapping $\hat{w} \rightarrow \hat{y}$. Yet when the initial state in (114) is zero, that mapping defines a Volterra operator of the second type, and is therefore invertible over any finite time span [24]. That is, $\mathcal{K}^\#$ and $\mathcal{K}_1^\#$ agree when restricted to the entire $L_2[s, t]$, for any finite s and t . Causality thus implies that $\mathcal{K}^\#$ and $\mathcal{K}_1^\#$ coincide throughout.

The proof is complete with the observation that (118) is just the conjugate of (116). \square

Proposition 3.5.2. *Suppose that Π_1 and Π_2 are as described in part (a) of Theorem 1, and that \mathcal{K} has a realization of the form (10). Then this compensator is internally stabilizing and strictly γ -suboptimal in (1). In particular, then $\gamma > \gamma_0$.*

Proof. The proof back-tracks the proof of necessity. We start by establishing that $\mathcal{K}^\#$, as given by (118), is internally stabilizing in the conjugate system (114).

Let the following be an exponentially stable realization of $\mathcal{H}_0^\#$:

$$(119) \quad \dot{\hat{r}} = M_0^\# \hat{r} + Q_0^\# \hat{q}, \quad \hat{v} = N_0^\# \hat{r} + R_0^\# \hat{q}.$$

We must show that when $\hat{w} = 0$ the combined closed-loop state of (114), (118), and (119),

$$\begin{bmatrix} \hat{x} \\ \hat{p} \\ \hat{r} \end{bmatrix},$$

satisfies an exponentially stable homogeneous ODE.

Obviously we can substitute $\hat{e} = \hat{x} - \hat{p}$ for \hat{p} in the extended state. As noted above, \hat{e} satisfies a homogeneous ODE; specifically,

$$(120) \quad \dot{\hat{e}} = A_1^\# \hat{e}.$$

Since A_1 is exponentially stable there is no loss of generality in taking $\hat{e} = 0$. We assume that to be the case, and focus on $\begin{bmatrix} \hat{x} \\ \hat{r} \end{bmatrix}$.

Our plan is to use Lemma 2.2. It thus must be checked (i) that the evolution of $\begin{bmatrix} \hat{x} \\ \hat{r} \end{bmatrix}$ is governed by a bounded perturbation of an exponentially stable generator, and (ii) that the norms

$$\left\| \begin{bmatrix} \hat{x} \\ \hat{r} \end{bmatrix} \right\|_{L_2[s, -t_0]}$$

are uniformly bounded in terms of the initial state $\begin{bmatrix} \hat{x}(s) \\ \hat{r}(s) \end{bmatrix}$.

We recall the observation made in the previous proof, that when $\hat{e} = 0$ and $\hat{w} = 0$ then $\hat{w}^\Delta = \hat{q}$. Thus (i) is easy. Our evolution is generated by

$$(121) \quad \begin{bmatrix} (A - (1/\gamma^2) B_1 B_1' \Pi_1 + \Pi_2 C_2' C_2)^\# - (1/\gamma^2) C_2^\# R_0^\# B_2^\# \Pi_1^\# \Pi_2^\# & C_2^\# N_0^\# \\ -(1/\gamma^2) Q_0^\# B_2^\# \Pi_1^\# \Pi_2^\# & M_0^\# \end{bmatrix}$$

¹² The objective of getting this equality was, in fact, the starting point in the derivation of (118) from (114).

By Proposition 3.3.1 (applied to (114)) the generator $(A - (1/\gamma^2)B_2B_2'\Pi_1 + \Pi_2C_1'C_1)^\#$ is exponentially stable. By assumption, so is $M_0^\#$. Thus our generator is indeed a bounded perturbation of an exponentially stable generator in the product space.

The following is the counterpart of (87) in the framework of (114)

$$(122) \quad \gamma^2 \|\hat{w}\|_{L_2[s,t]}^2 - \|\hat{z}\|_{L_2[s,t]}^2 = \langle \hat{x}(s), \Pi_2^\#(s)\hat{x}(s) \rangle - \langle \hat{x}(t), \Pi_2^\#(t)\hat{x}(t) \rangle \\ + \gamma^2 \|\hat{w}^\Delta\|_{L_2[s,t]}^2 - \|\hat{u}^\Delta\|_{L_2[s,t]}^2.$$

Taking $\hat{w} = 0$ and $t = -t_0$ (which is the maximal value for t in the conjugate system) and invoking assumption (7), equation (122) yields

$$(123) \quad \varepsilon \|\hat{x}\|_{L_2[s,-t_0]}^2 \leq \|\hat{z}\|_{L_2[s,-t_0]}^2 \\ = -\langle \hat{x}(s), \Pi_2^\#(s)\hat{x}(s) \rangle - \gamma^2 \|\hat{w}^\Delta\|_{L_2[s,-t_0]}^2 + \|\hat{u}^\Delta\|_{L_2[s,-t_0]}^2.$$

In adherence with previous notations pertinent to (1), we denote by $\hat{\mathcal{H}}(s): (r(s), y) \rightarrow r: R \times L_2(s, -t_0) \rightarrow L_2(s, -t_0)$ the mappings from the initial state and the input to the output in (119). (Thus $\mathcal{H}_0^\#|_{L_2(s,-t_0)} = \hat{\mathcal{H}}(s)|_{r(s)=0}$.) Due to the assumption that (119) is stable, these mappings are uniformly bounded. Let λ be a finite uniform bound on the induced norms $\|\hat{\mathcal{H}}(s)\|$. Set $\hat{e} = 0$ and $\hat{w} = 0$. Then there holds:

$$(124) \quad \|\hat{u}^\Delta\|_{L_2(s,-t_0)}^2 = \|\hat{\mathcal{H}}(s)(\hat{r}(s), \hat{w}^\Delta)\|_{L_2(s,-t_0)}^2 \\ = \|\hat{\mathcal{H}}(s)(0, \hat{w}^\Delta) + \hat{\mathcal{H}}(s)(\hat{r}(s), 0)\|_{L_2(s,-t_0)}^2 \\ \leq (\|\mathcal{H}_0^\# \hat{w}^\Delta\|_{L_2(s,-t_0)} + \|\hat{\mathcal{H}}(s)(\hat{r}(s), 0)\|_{L_2(s,-t_0)})^2 \\ \leq \gamma^2 \|\hat{w}^\Delta\|_{L_2(s,-t_0)}^2 + 2\gamma\lambda \|\hat{w}^\Delta\|_{L_2(s,-t_0)} \|\hat{r}(s)\| + \lambda^2 \|\hat{r}(s)\|^2.$$

Combining (123) and (124) we get

$$(125) \quad \varepsilon \|\hat{x}\|_{L_2[s,-t_0]}^2 \leq \|\Pi_2^\#\|_\infty \|\hat{x}(s)\|^2 + 2\gamma\lambda \|\hat{w}^\Delta\|_{L_2[s,-t_0]} \|\hat{r}(s)\| + \lambda^2 \|\hat{r}(s)\|^2 \\ = \|\Pi_2^\#\|_\infty \|\hat{x}(s)\|^2 + 2\frac{\lambda}{\gamma} \|B_1^{\#'} \Pi_1^\# \Pi_2^\# \hat{x}\|_{L_2[s,-t_0]} \|\hat{r}(s)\| + \lambda^2 \|\hat{r}(s)\|^2 \\ \leq \|\Pi_2^\#\|_\infty \|\hat{x}(s)\|^2 + 2\frac{\lambda}{\gamma} \|B_1^{\#'} \Pi_1^\# \Pi_2^\#\|_\infty \|\hat{x}\|_{L_2[s,-t_0]} \|\hat{r}(s)\| + \lambda^2 \|\hat{r}(s)\|^2,$$

which can be rewritten in the form

$$(126) \quad \varepsilon \|\hat{x}\|_{L_2[s,-t_0]}^2 - \eta \|\hat{x}\|_{L_2[s,-t_0]} \|\hat{r}(s)\| - \kappa \|\hat{x}(s)\|^2 - \lambda^2 \|\hat{r}(s)\|^2 \leq 0.$$

This last inequality provides a uniform bound of the form

$$(127) \quad \|\hat{x}\|_{L_2[s,-t_0]} \leq \alpha \|\hat{x}(s)\| + \beta \|\hat{r}(s)\|.$$

In turn, internal stability of (119) and the inequality (127) provide this next bound:

$$(128) \quad \|\hat{r}\|_{L_2[s,-t_0]} \leq \mu \|\hat{w}^\Delta\| - \frac{1}{\gamma^2} B_2^{\#'} \Pi_1^\# \Pi_2^\# \hat{x}\|_{L_2[s,-t_0]} + \nu \|\hat{r}(s)\| \\ \vdots \\ \leq \theta \|\hat{x}(s)\| + \tau \|\hat{r}(s)\|$$

for appropriate constants μ , ν , θ , and τ .

Relying on (127) and (128), we are now able to invoke Lemma 2.2 and deduce that, indeed, the evolution of $[\hat{x}_f^\#]$ is exponentially stable. Consequently, $\mathcal{H}^\#$ is internally stabilizing in (114); hence so is \mathcal{H} in (95).

Next we have to establish that $\mathcal{K}^\#$ is strictly γ -suboptimal in (114). Substituting $t = -t_0$ and $\hat{x}(s) = 0$ in (122), and invoking our assumption $\|\mathcal{H}_0\| < \gamma$, we get

$$(129) \quad \gamma^2 \|\hat{w}\|_{L_2[s, -t_0]}^2 - \|\hat{z}\|_{L_2[s, -t_0]}^2 > \delta^2 \|\hat{w}^\Delta\|$$

for some appropriate $\delta \neq 0$. With the initial values $\hat{x}(s) = \hat{p}(s) = 0$, the closed-loop mapping $\hat{w}^\Delta \rightarrow \hat{w}$ is realized by a stable system,

$$(130) \quad \begin{aligned} \dot{\hat{x}} &= A_2^\# \hat{x} - \Pi_1^\# B_2^{\#'} \hat{w}^\Delta + C_2^\# \hat{u}^\Delta \\ &= A_2^\# \hat{x} + (C_2^\# \mathcal{H}_0^\# - \Pi_1^\# B_2^{\#'}) \hat{w}^\Delta, \\ \hat{w} &= \frac{1}{\gamma^2} B_2^\# \Pi_1^\# \Pi_2^\# \hat{x} + \hat{w}^\Delta. \end{aligned}$$

In particular, it is bounded, and the inequality in (129) holds also when $\delta^2 \|\hat{w}^\Delta\|$ is replaced by $\theta^2 \|\hat{w}\|$ for some suitable $\theta \neq 0$. By Observation 3.0.1 the feedback $\mathcal{K}^\#$ is strictly γ -suboptimal in (114). By Corollary 3.4.2, \mathcal{K} is internally stabilizing and strictly γ -suboptimal in (1). \square

Remark. The trick by which (118), hence (10), was derived from (114) is simple: we first substitute \hat{u}^Δ for \hat{u} as the control input. That is done by appropriate modification of the state equation. Using the third equation in (114), we can substitute a term in \hat{x} and \hat{y} for the term in \hat{w} , in the modified state equation. Substituting \hat{p} and \hat{v} for \hat{x} and \hat{u}^Δ , the state equation in (118) is obtained. Using once more the third equation in (114), we get an expression for \hat{w}^Δ in terms of \hat{x} and \hat{y} . Substituting \hat{p} for \hat{x} , the third equation in (118) is obtained. The relation $\hat{v} = \mathcal{H}_0^\# \hat{q}$ then follows from the definition of \mathcal{H}_0 . The output formula in (118) comes directly from (117).

This completes the proof of Theorem I.

3.6. The corollary: Outline of proofs.

3.6.1. Part (e). If $A \in L_\infty$ then all the evolutions discussed heretofore are strongly differentiable [20, p. 129]. The Riccati differential equations (11) and (12) are obtained by differentiating (8) and (9).

Let $A_0 \triangleq A + B_2 B_2' \Lambda$, and let Φ_0 be the evolution it generates, as above. By the present assumption there holds

$$(131) \quad \|\Phi_0(t, s)\xi\| \geq e^{\|A_0\|_\infty(s-t)} \|\xi\|, \quad t \geq s.$$

Thus, when $t_1 = +\infty$ we have

$$(132) \quad \begin{aligned} J^0(s; \xi, 0) &= - \int_s^{+\infty} \langle \Phi_0(t, s)\xi, (C_1' C_1 + \Lambda B_2 B_2' \Lambda)(s) \Phi_0(t, s)\xi \rangle dt \\ &\leq -\varepsilon \int_s^{+\infty} \|\Phi_0(t, s)\xi\|^2 dt \\ &\leq -\frac{\varepsilon_1}{2\|A_0\|_\infty} \|\xi\|^2. \end{aligned}$$

Since

$$(133) \quad J^0(s; \xi, 0) \geq J^*(s; \xi) = \langle \xi, \Pi_1(s)\xi \rangle,$$

we conclude that Π_1 is uniformly negative. When $t_0 = -\infty$ the same reasoning assures uniform negativity of Π_2 .

3.6.2. Periodic and time invariant systems. When $t_1 = +\infty$, periodicity of the system coefficients implies periodicity of the min-max problem (27), and hence of its unique solution. In particular, then Π_1 is periodic. When we also have $t_0 = -\infty$, the same holds for Π_2 . When the system is time invariant, the assumption that the induced norm bound $\|\mathcal{T}_{\mathcal{H}}\| < \gamma$ can be achieved in stable closed-loop over some infinite ray, say $[t, +\infty)$ or $(-\infty, t]$, implies that it can be achieved for *any* such ray (i.e., for any choice of t). Consequently, then γ is strictly suboptimal over the entire real line. The result on periodic systems then implies that time invariant solutions to the two Riccati equations exist; in (8), (9) these solutions are defined with $t_0 = -\infty$ and $t_1 = +\infty$.

3.6.3. The asymptotic system. Our assumption here is that $t_0 = -\infty$, $t_1 = +\infty$, and that there exist A^+ , B_i^+ , C_i^+ , D_{ij}^+ , A^- , B_i^- , C_i^- , and D_{ij}^- such that the coefficients in (1) tend (in operator norm) to their “+” counterparts as $t \rightarrow +\infty$ and to their “-” counterparts as $t \rightarrow -\infty$. (e.g., $\|A(t) - A^+\| \rightarrow 0$ as $t \rightarrow +\infty$).

It follows from standard *small gain* arguments¹³ [16, Thm. 3] that \mathcal{H} is internally stabilizing in (1) if and only if so it is in the “+” and “-” asymptotic systems. Continuous dependence of the operator norms on the system parameters implies that strict γ -suboptimality of \mathcal{H} in the asymptotic systems is equivalent to strict γ -suboptimality of \mathcal{H} in (1) over intervals $(-\infty, s)$ and $(t, +\infty)$ for s small enough and t large enough. Therefore if solutions exist to the Riccati equations (8) and (9), as described in the theorem, then such solutions exist also in the asymptotic systems. Moreover, since the optimal min-max value $J^*(s; \xi) = \langle \xi, \Pi_1(s)\xi \rangle$ converges uniformly to its “+” counterpart as $s \rightarrow +\infty$, the convergence $\|\Pi_1(s) - \Pi_1^+\| \rightarrow 0$ follows. Similarly, we conclude that $\|\Pi_2(s) - \Pi_2^-\| \rightarrow 0$ as $s \rightarrow -\infty$. \square

4. Getting rid of removable assumptions. The technical discussion heretofore was carried under the assumptions of (7) on the structure of our system. The sole purpose of these hypotheses was to simplify notation, and *they are made without any loss of generality*. For completeness we verify this point now, and review a straightforward procedure for the removal of these assumptions. It is noted that there are several simple and more elegant ways to relax parts of (7); e.g., uniform positivity of $C_1' C_1$ can be substituted by uniform observability of $[A, C_1]$. Several authors (e.g., [16], [25], [27]–[29]) explored such ways, especially for the finite-dimensional, LTI case; their papers provide interesting and insightful analysis, and bear potential advantages. Since (given more awkward notation and definitions) the logic of the proofs remains exactly as in the previous section, proofs will be omitted. We shall be content with highlights of the modifications in formulae, statements of main results, and of those few arguments that have to be added.

The hypotheses of (7) can be divided into subgroups: the separation assumptions $D_{12}' C_1 = 0$ and $D_{21} B_1' = 0$, the normalized input-nonsingularity assumptions $D_{12}' D_{12} = I$ and $D_{21} D_{21}' = I$, the state-nonsingularity assumptions $C_1' C_1 > \varepsilon I$ and $B_1 B_1' > \varepsilon I$, the zero-tracking objective assumption $D_{11} = 0$, and the strict closed-loop causality assumption $D_{22} = 0$. Other than the last two, these assumptions came in pairs: one applies to the optimal state-feedback problem, while the other applies to the conjugate, estimation problem. We shall focus mostly on the state-feedback part; the conjugate system is to be treated in complete analogy.

The first Riccati equation. The following simple statement will be used to overcome singularity with respect to the state and/or control.

¹³ These arguments state that small perturbations do not destroy exponential stability.

PROPOSITION 4.1. *Let γ_0 be the optimal value in (1). Then $\gamma > \gamma_0$ if and only if γ is a strictly suboptimal value when the system's output is modified to¹⁴*

$$(134) \quad z_n \triangleq \begin{bmatrix} z \\ \varepsilon_0 x \\ \varepsilon_0 u \end{bmatrix}$$

for some $\varepsilon_0 > 0$. In particular, once such γ and ε_0 are fixed, the same holds when ε_0 is substituted by any $\varepsilon \in [0, \varepsilon_0]$.

Proof. Suppose $\gamma > \gamma_0$ in (1). Then there exist $\delta \neq 0$ and an admissible internally stabilizing feedback compensator, $u = \mathcal{K}y$, so that

$$(135) \quad \|z\|^2 \leq (\gamma^2 - \delta^2) \|w\|^2$$

in closed-loop. Since the closed-loop mappings $w \rightarrow x, u$ are continuous, we can choose ε_0 small enough to guarantee the closed-loop inequality

$$(136) \quad \varepsilon_0^2 (\|x\|^2 + \|u\|^2) \leq \frac{1}{2} \delta^2 \|w\|^2$$

whereby

$$(137) \quad \|z_n\|^2 \leq (\gamma^2 - \frac{1}{2} \delta^2) \|w\|^2.$$

Thus γ is strictly suboptimal in the modified system. The converse claim is obvious. \square

In view of the observation we substitute z by z_n . The modified output satisfies $\|z_n\|^2 \geq \varepsilon^2 \|x\|^2, \varepsilon^2 \|u\|^2$.

The purpose of the next set of modifications is to rewrite $\|z_n\|$ in the form

$$(138) \quad \|z_n\|^2 = \|C_{n1}(\varepsilon)(x + E_0(\varepsilon)w)\|^2 + \|E_1(\varepsilon)w\|^2 \\ + \|D_{n12}(\varepsilon)(u_n + E_2(\varepsilon)w)\|^2.$$

That can be done, setting

$$(139) \quad \begin{aligned} L(\varepsilon) &\triangleq -(D'_{12}D_{12} + \varepsilon^2 I_U)^{-1} D'_{12} C_1, \\ u_n &\triangleq u - L(\varepsilon)x, \\ D_{n12}(\varepsilon) &\triangleq \begin{bmatrix} D_{12} \\ 0_X \\ \varepsilon I_U \end{bmatrix}, \\ C_{n1}(\varepsilon) &\triangleq \begin{bmatrix} C_1 + D_{12}L(\varepsilon) \\ \varepsilon I_X \\ \varepsilon L \end{bmatrix}, \\ E_0(\varepsilon) &\triangleq (C'_{n1}(\varepsilon)C_{n1}(\varepsilon))^{-1} C'_{n1}(\varepsilon) \begin{bmatrix} D_{11} \\ 0_X \\ 0_U \end{bmatrix}, \\ E_2(\varepsilon) &\triangleq (D'_{n12}(\varepsilon)D_{n12}(\varepsilon))^{-1} D'_{n12}(\varepsilon) \begin{bmatrix} D_{11} \\ 0_X \\ 0_U \end{bmatrix}, \\ E_1(\varepsilon) &\triangleq \begin{bmatrix} D_{11} \\ 0_X \\ 0_U \end{bmatrix} - C_{n1}(\varepsilon)E_0(\varepsilon) - D_{n12}(\varepsilon)E_2(\varepsilon). \end{aligned}$$

¹⁴ In what follows the subscript n stands for "new," where variables or coefficients will be modified.

For brevity we shall suppress the notation of ε , bearing in mind the dependence of the *new* variables on that constant.

Consider now the modified version of the min-max problem (27):

$$(140) \quad \min_{w \in L_2} \max_{u \in L_2} \{ \gamma^2 \|w\|^2 - \|z_n\|^2 \}.$$

Once admissible controls are restricted to be only those that result in L_2 output (hence L_2 -state trajectories), the class of admissible u 's stands in 1:1 correspondence to the class of admissible u_n 's. We thus can, and henceforth shall, substitute u_n for u in (140), which thereby becomes

$$(141) \quad \min_{w \in L_2} \left\{ (\gamma^2 \|w\|^2 - \|E_1 w\|^2) - \min_{u_n \in L_2} \{ \|C_{n1}(x + E_0 w)\|^2 + \|D_{n12}(u_n + E_2 w)\|^2 \} \right\}.$$

It is a tracking problem, where the target trajectories for both x and u_n are determined by w .

The solution of the modified min-max problem goes exactly as developed in §§ 3.1 and 3.2 (given the cumbersome burden of more complicated variables and coefficients). In particular, the nonsingularity of the cost functional and strict suboptimality of γ guarantee existence, uniqueness, and continuous dependence on the initial data of a solution to the following Hamilton-Jacobi system:

$$(142) \quad \begin{aligned} \dot{x} &= (A + (B_1 - B_2 E_2)(S_1 - E'_0 S_0 E_0)^{-1} E'_0 S_0 + B_2 L)x \\ &\quad + (B_2 S_2^{-1} B'_2 - (B_1 - B_2 E_2)(S_1 - E'_0 S_0 E_0)^{-1} (B_1 - B_2 E_2)') e \\ \dot{e} &= (S_0 + S_0 E_0 (S_1 - E'_0 S_0 E_0)^{-1} E'_0 S_0) x \\ &\quad - (A + (B_1 - B_2 E_2)(S_1 - E'_0 S_0 E_0)^{-1} E'_0 S_0 + B_2 L)' e, \end{aligned}$$

where we use the abbreviations

$$(143) \quad \begin{aligned} S_0 &\triangleq C'_{n1} C_{n1}, \\ S_1 &\triangleq \gamma^2 I - E'_1 E_1, \\ S_2 &\triangleq D'_{n12} D_{n12}. \end{aligned}$$

When considered over an interval (s, t_1) , this system is subject to the boundary conditions $x(s) = \xi$ and $e(t_1) = 0$ (or $\lim_{t \rightarrow +\infty} e(t) = 0$ when $t_1 = +\infty$). The min-max inputs are the given by

$$(144) \quad \begin{aligned} w^* &= (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 x^* - (B_1 - B_2 E_2)' e^*), \\ u^* &= Lx^* + S_2^{-1} B'_2 e^* - E_2 w^*. \end{aligned}$$

Exactly as in the proof of uniqueness of the solution of the Hamilton-Jacobi systems (35) and (73), by (43), we use the following facts in establishing uniqueness of the solution of (143): (i) A solution to (143) defines an optimal solution to the min-max problem (142); (ii) the optimal state and input trajectories in the latter are unique; (iii) when we set $x^* = 0$, $u^* = 0$, and $w^* = 0$ then the e component of the solution satisfies a homogeneous ODE, now $\dot{e} = -(A + B_2 L)' e$, the zero terminal condition, and $B'_2 e = 0$; finally, (iv) the pair $[A, B_2]$, and hence $[A + B_2 L, B_2]$, is stabilizable.

The associated Riccati equation is

$$(145) \quad \begin{aligned} \Pi_1(t) = & - \int_t^{t_1} \Phi_1(s, t)' ((S_0 + S_0 E_0 (S_1 - E_0' S_0 E_0)^{-1} E_0' S_0) \\ & + \Pi_1(B_2 S_2^{-1} B_2' - (B_1 - B_2 E_2)(S_1 - E_0' S_0 E_0)^{-1} \\ & \cdot (B_1 - B_2 E_2)' \Pi_1)(s) \Phi_1(s, t) ds, \end{aligned}$$

where Φ_1 is the evolution generated by

$$(146) \quad \begin{aligned} A_1 \triangleq & A + (B_1 - B_2 E_2)(S_1 - E_0' S_0 E_0)^{-1} E_0' S_0 + B_2 L \\ & + (B_2 S_2^{-1} B_2' - (B_1 - B_2 E_2)(S_1 - E_0' S_0 E_0)^{-1} (B_1 - B_2 E_2)') \Pi_1. \end{aligned}$$

(We easily see that when (7) holds and $\varepsilon = 0$, this equation and generator retain their form from the previous sections.)

The appropriate counterpart of the complete-observation result, Theorem II, is given in terms of this Riccati equation as follows.

THEOREM 4.2. *Suppose that both the state x and (for small ε) the control-target trajectory $E_2 w$ are available. Then γ is strictly suboptimal if and only if the operator $S_1 - E_0' S_0 E_0$ is uniformly positive and there exists a uniformly bounded, negative-definite solution to the Riccati equation (145) with the associated A_1 being an exponentially stable generator, for some choice of small, nonzero ε . Moreover, if indeed such a solution exists for one choice of ε , it also exists for any smaller ε , the state-feedback control $u = (L + (D_{n12}' D_{n12})^{-1} B_2')x$ is internally stabilizing, and the feedback-feedforward control $u = (L + (D_{n12}' D_{n12})^{-1} B_2')x - E_2 w$ guarantees the closed-loop norm bound $\|\mathcal{T}_{\mathcal{K}}\| < \gamma$.*

The proof of sufficiency is essentially the same as in Theorem II. The following justifies the added constraint in the necessary part, the uniform positivity of $S_1 - E_0' S_0 E_0$.

PROPOSITION 4.3. *If (137) holds in stable closed-loop (for some positive ε and δ) then $S_1 - E_0' S_0 E_0$ is uniformly positive definite.*

Outline of the proof. The stable closed-loop mapping $w \rightarrow x$ is strictly causal. Thus, over short periods of time, its induced norm can be made arbitrarily small. The inequality (137) should hold over any time interval, and if the interval is sufficiently short, it thus implies

$$(147) \quad \begin{aligned} \frac{1}{4} \delta^2 \|w\|_{L_2[s, t]}^2 & \leq \gamma^2 \|w\|_{L_2[s, t]}^2 - \|C_{n1} E_0 w\|_{L_2[s, t]}^2 - \|E_1 w\|_{L_2[s, t]}^2 \\ & \quad - \|D_{n12}(u + E_2 w)\|_{L_2[s, t]}^2 \\ & \leq \gamma^2 \|w\|_{L_2[s, t]}^2 - \|C_{n1} E_0 w\|_{L_2[s, t]}^2 - \|E_1 w\|_{L_2[s, t]}^2 \\ & = \langle w, (S_1 - E_0' S_0 E_0) w \rangle_{L_2[s, t]}. \end{aligned}$$

Consequently,

$$(148) \quad \frac{1}{4} \delta^2 \|w(t)\|_{\mathbf{W}}^2 \leq \langle w(t), (\gamma^2 I - E_0' C_{n1}' C_{n1} E_0 - E_1' E_1)(t) w(t) \rangle_{\mathbf{W}}$$

almost everywhere. \square

Conjugation and the second Riccati equation. As in the case where (7) holds, the proof of Theorem 4.2 and the entire ensuing analysis, rely heavily on the following modified version of (87):¹⁵

$$(149) \quad \begin{aligned} & \|w\|_{S_1}^2 - \|x + E_0 w\|_{S_0}^2 - \|u - E_2 w\|_{S_2}^2 \\ & = \|\xi\|_{\Pi_1}^2 + \|w^\Delta\|_{(S_1 - E_0' S_0 E_0)}^2 - \|u^\Delta\|_{S_2}^2 \end{aligned}$$

¹⁵ Here the norm-notation form $\|\bullet\|_{\Theta}$ stands for the Hilbert-space norm weighted by Θ : $\|\psi\|_{\Theta}^2 \triangleq \langle \psi, \Theta \psi \rangle$.

where we use the modified definitions for the momentary deviations from optimality

$$\begin{aligned}
 w^\Delta &\triangleq w - (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1)_x \\
 u^\Delta &\triangleq u + E_2 w - (L + S_2^{-1} B_2' \Pi_1) x \\
 &= u + E_2 w^\Delta - (L + S_2^{-1} B_2' \Pi_1 - E_2 (S_1 - E'_0 S_0 E_0)^{-1} \\
 &\quad \cdot (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1)) x.
 \end{aligned}
 \tag{150}$$

The estimation problem is based on the conjugate of the closed-loop mapping $w^\Delta \rightarrow u^\Delta$. Given in (1) an admissible, internally stabilizing feedback control $u = \mathcal{K}y$ that maintains (137), that mapping is realized by the following system:

$$\begin{aligned}
 \dot{x} &= (A + B_1 (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1)) x \\
 &\quad + B_1 w^\Delta + B_2 u \\
 u^\Delta &= -(L + S_2^{-1} B_2' \Pi_1 - E_2 (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1)) x \\
 &\quad + E_2 w^\Delta + u \\
 y &= (C_2 + D_{21} (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1)) x \\
 &\quad + D_{21} w^\Delta + D_{22} u
 \end{aligned}
 \tag{151}$$

subject to $\hat{u} = \mathcal{H}y$; by Proposition 3.4.1, that feedback is internally stabilizing and strictly γ -suboptimal in (151). The conjugate mapping is thus realized by the well-defined, internally stable, closed-loop system

$$\begin{aligned}
 \dot{\hat{x}} &= (A + B_1 (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1))^\# \hat{x} \\
 &\quad - (L + S_2^{-1} B_2' \Pi_1 - E_2 (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1))^\# \hat{w} \\
 &\quad + (C_2 + D_{21} (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1))^\# \hat{u} \\
 \hat{z} &= B_1^\# \hat{x} + E_2^\# \hat{w} + D_{21}^\# \hat{u}, \\
 \hat{y} &= B_2^\# \hat{x} + \hat{w} + D_{22}^\# \hat{u}, \quad \hat{u} = \mathcal{H}^\# \hat{y}.
 \end{aligned}
 \tag{152}$$

Now we repeat our previous constructions: Provided that ε is small enough, we can substitute the output of (152) by \hat{z}_n such that $\|\hat{z}_n\|^2 = \|\hat{z}\|^2 + \varepsilon^2 (\|\hat{x}\|^2 + \|\hat{u}\|^2)$, without affecting the strict γ -suboptimality of the compensator. Then just as before, we introduce a control change of variable $\hat{u}_n \triangleq \hat{u} - M^\# \hat{x}$, and new and modified coefficients $B_{n1}^\#(\varepsilon)$, $D_{n21}^\#(\varepsilon)$, $F_0^\#(\varepsilon)$, $F_1^\#(\varepsilon)$, and $F_2^\#(\varepsilon)$, in terms of which we further denote $T_0^\# \triangleq B_{n1}^\# B_{n1}^\#$, $T_1^\# \triangleq \gamma^2 I - F_1^\# F_1^\#$, and $T_2^\# \triangleq D_{n21}^\# D_{n21}^\#$, so that

$$\|\hat{z}_n\|^2 = \|B_{n1}^\# (\hat{x} + F_0^\# \hat{w})\|^2 + \|F_1^\# \hat{w}\|^2 + \|D_{n21}^\# (\hat{u}_n + F_2^\# \hat{w})\|^2
 \tag{153}$$

and

$$\gamma^2 \|\hat{w}\|^2 - \|\hat{z}_n\|^2 = \|\hat{w}\|_{T_0^\#}^2 - \|(\hat{x} + F_0^\# \hat{w})\|_{T_0^\#}^2 (\hat{u}_n + F_2^\# \hat{w})\|_{T_2^\#}^2.
 \tag{154}$$

In order that the estimation Riccati equation, which involves more terms than its control counterpart, (145), be written in a readable way, we make three more abbreviations:

$$\begin{aligned}
 \tilde{A}^\# &\triangleq (A + B_1 (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1))^\# \\
 G_1^\# &\triangleq -(L + S_2^{-1} B_2' \Pi_1 - E_2 (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1))^\# \\
 &\quad - (C_2 + D_{21} (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1))^\# F_2^\#, \\
 G_2^\# &\triangleq (C_2 + D_{21} (S_1 - E'_0 S_0 E_0)^{-1} (E'_0 S_0 - (B_1 - B_2 E_2)' \Pi_1))^\#.
 \end{aligned}
 \tag{155}$$

In these terms, the second Riccati equation is

$$(156) \quad \begin{aligned} \Pi_2(s) = & - \int_{t_0}^s \Phi_2(s, t) ((T_0 + T_0 F'_0 (T_1 - F_0 T_0 F'_0)^{-1} F_0 T_0) \\ & + \Pi_2(G'_2 T_2^{-1} G_2 - G'_1 (T_1 - F_0 T_0 F'_0)^{-1} G_1) \Pi_2)(t) \Phi_2(s, t)' dt \end{aligned}$$

where Φ_2 is generated by

$$(157) \quad \begin{aligned} A_2 \triangleq & \tilde{A} + M G_2 + T_0 F'_0 (T_1 - F_0 T_0 F'_0)^{-1} G_2 \\ & + \Pi_2(G'_2 T_2^{-1} G_2 - G'_1 (T_1 - F_0 T_0 F'_0)^{-1} G_1). \end{aligned}$$

The conclusion at this point is, as expected, the following proposition.

Proposition 4.4. *If γ is strictly suboptimal in (1) then there exist uniformly bounded, negative-definite solutions to the two Riccati equations (145) and (156), with A_1 and A_2 being exponentially stable generators, for some choice of ε . Moreover, if such solutions exist for $\varepsilon = \varepsilon_0$ then the same holds for all $\varepsilon \in (0, \varepsilon_0)$.*

Parametrization of all compensators. We start with the case of $D_{22} = 0$. The momentary deviation of the control and the disturbance of (153) from their optimal values are given by

$$(158) \quad \begin{aligned} \hat{w}^\Delta & \triangleq \hat{w} - (T_1 - F_0 T_0 F'_0)^{\#-1} (T_0 F'_0 - \Pi_2 G'_1)^{\#} \hat{x}, \\ \hat{u}^\Delta & \triangleq \hat{u} + F_2^{\#} \hat{w} - (M + \Pi_2 G'_2 T_2^{-1})^{\#} \hat{x}. \end{aligned}$$

Following the directions in the remark at the end of § 3.5 (adapted from (114) to (152) with $D_{22} = 0$), we use these definitions to obtain this next representation for $\mathcal{K}^{\#}$:

$$(159) \quad \begin{aligned} \dot{\hat{p}} & = (\tilde{A} - B_2 G_1 + (M + \Pi_2 G'_2 T_2^{-1} + B_2 F_2) G_2)^{\#} \hat{p} + G_1^{\#} \hat{y} + G_2^{\#} \hat{v}, \\ \hat{u} & = (M + \Pi_2 G'_2 T_2^{-1} + B_2 F_2)^{\#} \hat{p} - F_2^{\#} \hat{y} + \hat{v}, \\ \hat{q} & = -(B^2 + (T_0 F'_0 - \Pi_2 G'_1)(T_1 - F_0 T_0 F'_0)^{-1})^{\#} \hat{p} + \hat{y}, \\ \hat{v} & = \mathcal{K}_0^{\#} \hat{q}. \end{aligned}$$

This system's conjugate provides the parametrization of compensators in (1), as follows.

Proposition 4.5. *Suppose that $D_{22} = 0$. Then (a) γ is strictly suboptimal in (1) if and only if for some small ε there exist uniformly bounded, negative-definite solutions to the Riccati equations (145) and (156), such that the associated generators A_1 and A_2 are exponentially stable. Moreover, if such solutions exist for $\varepsilon = \varepsilon_0$ then the same holds for all $\varepsilon \in (0, \varepsilon_0)$. (b) If γ is strictly suboptimal in (1) then admissible, internally stabilizing, strictly γ -suboptimal compensators in (1) are those having realizations of the following form:*

$$(160) \quad \begin{aligned} \dot{p} & = (\tilde{A} - B_2 G_1 + (M + \Pi_2 G'_2 T_2^{-1} + B_2 F_2) G_2) p \\ & + (M + \Pi_2 G'_2 T_2^{-1} + B_2 F_2) y \\ & - (B_2 + (T_0 F'_0 - \Pi_2 G'_1)(T_1 - F_0 T_0 F'_0)^{-1}) v \\ u & = G_1 p - F_2 y + v \\ q & = G_2 p + y, \quad v = \mathcal{K}_0 q \end{aligned}$$

where \mathcal{K}_0 is the input-output mapping in an exponentially stable system, $\|\mathcal{K}_0\| < \gamma$, and the solutions to the Riccati equations are derived for some small enough ε .

The arguments in the proofs of Propositions 3.5.1 and 3.5.2, which are needed in order to complete the proof of the current statement (beyond Proposition 4.4), adapt with none but notational changes to the current situation.

When treating the general case, $D_{22} \neq 0$, we recall the following simple fact.

OBSERVATION 4.6. *Let*

$$(161) \quad \dot{x} = Ax + Bu, \quad y = Cs + Du$$

be a given system, satisfying our standard assumptions. Then the following is an internally stabilizing compensator when we set $D = 0$,

$$(162) \quad \dot{p} = Mp + Ny, \quad u = Qp + Ry,$$

and $(I + RD)^{-1} \in L_\infty$, if and only if this next compensator is admissible and internally stabilizing in the original setting.

$$(163) \quad \begin{aligned} \dot{p} &= (M - N(I + DR)^{-1}DQ)p + N(I + DR)^{-1}y, \\ u &= (I + RD)^{-1}Qp + (I + RD)^{-1}Ry. \end{aligned}$$

Outline of the proof. One easily notes that, provided $(I + RD)^{-1} \in L_\infty$, the substitution of $y - Du$ for y in (162) yields (163). That is, the same closed-loop state $\begin{pmatrix} x \\ p \end{pmatrix}$ will satisfy both (161), (162) and (161), (163) for the two alternative output functions. In particular, (162) is internally stabilizing when $D = 0$ if and only if (163) is internally stabilizing in the original setting. The equality $(I - (I + RD)^{-1}RD)^{-1} = (I + RD)^{-1}$ shows that, indeed, $(I + RD)^{-1} \in L_\infty$ is the appropriate admissibility condition. \square

Direct application of this observation to proposition 4.5 yields the following general result.

THEOREM 4.7. (a) γ is strictly suboptimal in (1) if and only if for some ε small enough the operator $S_1 - E'_0 S_0 E_0$ is uniformly positive, there exists a uniformly bounded, negative-definite solution to the Riccati equation (145) such that A_1 is exponentially stable, the operator $T_1 - F_0 T_0 F'_0$ is uniformly positive, and the Riccati equation (156) possesses a uniformly bounded, negative-definite solution such that the associated generators and A_2 is exponentially stable. Moreover, if such solutions exist for $\varepsilon = \varepsilon_0$ then the same holds for all $\varepsilon \in (0, \varepsilon_0)$. (b) If γ is strictly suboptimal in (1) then admissible, internally stabilizing, strictly γ -suboptimal compensators in (1) are those having realizations as follows: the design-free parameter is a stable linear system \mathcal{H}_0 ,

$$(164) \quad \dot{\phi} = M_0 \phi + N_0 y, \quad \psi = Q_0 \phi + R_0 y$$

with the properties (i) $\|\mathcal{H}_0\| < \gamma$ and (ii) $(I + (R_0 - F_2)D_{22})^{-1} \in L_\infty$; the compensator associated with \mathcal{H}_0 is then of the form

$$(165) \quad \begin{aligned} \begin{pmatrix} \dot{p} \\ r \end{pmatrix} &= \left(\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} - \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} (I + D_{22}(R_0 - F_2))^{-1} [\Theta_1 & \Theta_2] \right) \begin{pmatrix} p \\ r \end{pmatrix} \\ &+ \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} (I + D_{22}(R_0 - F_2))^{-1} y, \\ u &= (I + (R_0 - F_2)D_{22})^{-1} [\Theta_1 & \Theta_2] \begin{pmatrix} p \\ r \end{pmatrix} + (r_0 - F_2)y \end{aligned}$$

where

$$\begin{aligned} \Gamma_{11} &\triangleq (\tilde{A} - B_2 G_1 + (M + \Pi_2 G'_2 T_2^{-1} + B_2 F_2) G_2) \\ &\quad - (B_2 + (T_0 F'_0 - \Pi_2 G'_1)(T_1 - F_0 T_0 F'_0)^{-1}) R_0 G_2 \\ \Gamma_{12} &\triangleq -(B_2 + (T_0 F'_0 - \Pi_2 G'_1)(T_1 - F_0 T_0 F'_0)^{-1}) Q_0 \\ \Gamma_{21} &\triangleq N_0 G_2 \end{aligned}$$

$$\begin{aligned}
 (166) \quad & \Gamma_{22} \triangleq M_0 \\
 & \Delta_1 \triangleq (M + \Pi_2 G_2' T_2^{-1} + B_2 F_2) \\
 & \quad - (B_2 + (T_0 F_0' - \Pi_2 G_1')(T_1 - F_0 T_0 F_0')^{-1}) R_0 \\
 & \Delta_2 \triangleq N_0 \\
 & \Theta_1 \triangleq G_1 + R_0 G_2 \\
 & \Theta_2 \triangleq Q_0
 \end{aligned}$$

and where the solutions to the Riccati equations are derived for some small ϵ .

Acknowledgment. The author thanks Ruth Curtain and two anonymous referees for their critical reading and useful comments.

REFERENCES

- [1] B. D. O. ANDERSON AND J. B. MOORE, *Optimal Control: Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1990.
- [2] T. BASAR, *A dynamic games approach to control design: disturbance rejection in discrete time*, Proc. 28th CDC, Tampa, FL, 1989, pp. 407-414.
- [3] J. BALL AND N. COHEN, *Sensitivity minimization in an H_∞ norm: parametrization of all suboptimal solutions*, Internat. J. Control, 46 (1987), pp. 785-816.
- [4] D. S. BERNSTEIN AND W. M. HADDAD, *LQG control with an H_∞ performance bound: a Riccati equation approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 293-305.
- [5] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, Berlin, 1978.
- [6] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite Riccati equations for systems defined by evolution operators*, SIAM J. Control Optim., 14 (1976), pp. 951-983.
- [7] R. DATKO, *A linear control problem in abstract Hilbert spaces*, J. Differential Equations, 9 (1971), pp. 346-359.
- [8] ———, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428-445.
- [9] J. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831-847.
- [10] B. F. FRANCIS, *A Course in H_∞ Control Theory*, Lecture Notes in Control Inform. Sci. 88, Springer-Verlag, Berlin, 1987.
- [11] J. S. GIBSON, *The Riccati integral equation for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 14 (1976), pp. 537-565.
- [12] M. J. GRIMBLE, *Optimal H_∞ robustness and the relationship to LQG design problems*, Internat. J. Control, 43 (1986), pp. 351-372.
- [13] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk sensitivity*, System Control Lett., 11 (1988), pp. 167-172.
- [14] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
- [15] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTE, *H_∞ optimal control and state feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 786-788.
- [16] P. P. KHARGONEKAR, I. R. PETERSEN, AND K. ZHOU, *Robust stabilization of uncertain systems: Quadratic stabilization and H_∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356-361.
- [17] D. J. LIMEBEER, B. D. O. ANDERSON, P. P. KHARGONEKAR, AND M. GREEN, *A game theoretic approach to H_∞ control for time varying systems*, To appear in SIAM J. Control and Optim.
- [18] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [19] D. MUSTAFA, *Relations between maximum entropy/ H_∞ control and combined H_∞ /LQG control*, System Control Lett., 12 (1989), pp. 193-203.
- [20] A. PAZY, *Semigroups of Linear Operators and Relations to Differential Equations*, Appl. Math. Sci., 44, Springer-Verlag, New York, 1983.
- [21] I. R. PETERSEN, *Linear quadratic differential games with cheap control*, System Control Lett., 8 (1986), pp. 181-188.

- [22] ———, *Stabilization of an uncertain linear systems in which uncertain parameters enter into the input matrix*, SIAM J. Control Optim., 26 (1988), pp. 1257–1264.
- [23] M. A. ROTEA AND P. P. KHARGONEKAR, *Stabilization of linear time varying and uncertain linear systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 884–887.
- [24] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Frederick Unger, New York, 1955.
- [25] R. RAVI, K. N. NAGPAL, AND P. P. KHARGONEKAR, *Control of linear time-varying systems: a state space approach*, SIAM J. Control and Optim., 29 (1991), pp. 1394–1413.
- [26] I. RHEE AND J. L. SPEYER, *A game theoretic controller and its relationship to H_∞ and linear-exponential-gaussian synthesis*, Proc. 28th CDC, Tampa, FL, 1989, pp. 909–915.
- [27] A. A. STOORVOGEL, *The H_∞ control problem: A state space approach*, Ph.D. thesis, Eindhoven, July 1990.
- [28] M. G. SAFONOV, D. J. N. LIMEBEER, AND R. Y. CHIANG, *Simplifying the H_∞ theory via loop-shifting, matrix-pencil and descriptor concepts*, Internat. J. Control, 50 (1989), pp. 2467–2488.
- [29] A. A. STOORVOGEL AND H. L. TRENTELMAN, *The quadratic matrix inequality in singular H_∞ control with state feedback*, SIAM J. on Control and Optim., to appear.
- [30] G. TADMOR, *H_∞ in the time domain: the standard four block problem*, LCDS/CSS Tech. Report #88-21, July 1988, Brown University, Providence, RI.
- [31] ———, *Worst case design in the time domain: the maximum principle and the standard H_∞ problem*, MCSS, 3 (1990), pp. 301–324.
- [32] ———, *I/O norms in general linear systems*, Internat. J. Control 51 (1990), 911–921.
- [33] ———, *H_∞ optimal sampled-data control in continuous time system*, Internat. J. Control, 56 (1992), pp. 99–141.
- [34] ———, *Receding horizon revisited: An easy way to robustly stabilize an LTV system*, System Control Lett., 18 (1992), pp. 285–294.
- [35] ———, *Uncertain feedback loops and robustness in general linear system*, Automatica, 27 (1991), pp. 1039–1042.
- [36] ———, *Time domain optimal control and worst case linear system design*, Proc. 28th CDC, Tampa, FL, 1989, pp. 403–406.
- [37] ———, *H_∞ in the time domain: the standard four block problem*, Proc. 1989 ACC, pp. 772–773.
- [38] M. S. VERMA AND J. C. ROMIG, *Reduced order controller in H_∞ -optimal synthesis methods of the first kind*, Math. Systems Theory, 22 (1989), pp. 109–148.
- [39] I. YAESH AND U. SHAKED, *Minimum H_∞ -norm regulation of linear discrete-time systems and its relation to linear quadratic discrete game*, Proc. 28th CDC, Tampa, FL, 1989, pp. 942–947.
- [40] K. ZHOU AND P. P. KHARGONEKAR, *An algebraic Riccati equation approach to H_∞ optimization*, System Control Lett., 11 (1988), pp. 85–92.

APPROXIMATION OF THE ALGEBRAIC RICCATI EQUATION IN THE HILBERT SPACE OF HILBERT–SCHMIDT OPERATORS*

A. DE SANTIS†, A. GERMANI‡, AND L. JETTO§

Abstract. This paper deals with the problem of approximating the infinite-dimensional algebraic Riccati equation, considered as an abstract equation in the Hilbert space of Hilbert–Schmidt operators. Two kinds of approximating schemes are proposed. The first scheme exploits the already established approximability of the corresponding dynamical Riccati equation together with its time convergence toward the steady state. The second method considers a particular sequence of finite-dimensional linear equations whose solutions are proved to converge toward the exact steady-state solution of the original problem.

Key words. infinite-dimensional systems, Galerkin approximation, algebraic Riccati equation

AMS subject classifications. Q3C25, Q3E11, 41A65, 65L60

1. Introduction. Both linear quadratic (LQ) optimal control and optimal linear filtering problems for linear systems evolving in Hilbert spaces lead to an infinite-dimensional Riccati equation. This has motivated the wide interest that, for at least two decades, has been devoted to establishing conditions for the existence and uniqueness of the solution of this equation [6], [9], [10], [12]. This problem has also been considered in [8], [15], [16], [20], [26], and [32] with particular reference to the LQ optimal control, in [6], [7], [19], [29], and [36] with reference to the optimal linear filtering, and in [10] and [11] with reference to both cases. In the above papers the topic is treated in different settings according to the different forms that the infinite-dimensional Riccati equation can take, depending on the structure assumed for the system dynamics.

Because of the infinite dimensionality, this Riccati equation cannot be instrumented by the standard computation techniques used in the classical finite-dimensional case and requires various approximation methods and truncation techniques. A large number of papers have been devoted to this problem. See, for instance, [4], [14], [22], [24], [25], [27], and [28]. The methods described in these papers work by projecting the infinite-dimensional Riccati equation onto a sequence of finite-dimensional subspaces of the original Hilbert space. The exact solution of the actual Riccati equation is so approximated by a sequence of solutions of finite-dimensional approximate Riccati equations.

Particularly important is the so-called *infinite-horizon problem*, which arises when the Riccati equation admits a steady-state solution. The corresponding nondynamical equation is referred to as the *algebraic Riccati equation* (ARE) [11], [18], [20], and [38]. Unfortunately, in this case the approximation problem tends to be much more difficult than the finite-horizon problem, for which significant results are available. Such a problem was discussed in [3], [21], and [22] with reference to the LQ optimal control, under the hypothesis that both the approximate semigroups and their adjoints

* Received by the editors June 4, 1990; accepted for publication (in revised form) December 3, 1991.

† Istituto di Analisi dei Sistemi ed Informatica del Consiglio Nazionale delle Ricerche, Viale Manzoni 30, 00185 Roma, Italy.

‡ Dipartimento di Ingegneria Elettrica, Università dell'Aquila, 67100 Monteluco (L'Aquila), Italy, and Istituto di Analisi dei Sistemi ed Informatica del Consiglio Nazionale delle Ricerche, Viale Manzoni 30, 00185 Roma, Italy.

§ Dipartimento di Elettronica ed Automatica, Università di Ancona, via Breccie Bianche, 67100 Ancona, Italy.

converge in the Trotter-Kato sense. Moreover, these papers also assumed uniform stability of the approximating semigroups.

The finite-dimensional approximation of the ARE in the space of Hilbert-Schmidt (H.S.) operators has recently been investigated in [33] and [34] under the hypothesis that the generator of the semigroup governing the system is strongly coercive.

In the present paper the approximation problem of the ARE in Hilbert spaces is considered with reference to the optimal-linear-filtering problem. As it is well acknowledged (see, e.g., [1], [6], and [7]), such a problem can be formally stated as follows.

Let us consider the following linear infinite-dimensional system on the Hilbert space H :

$$(1.1) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + B\omega(t), & x(0) &= x_0, \\ y(t) &= Cx(t) + G\omega(t), \end{aligned}$$

where the following hold:

(a) A is the infinitesimal generator of a strongly continuous semigroup $\{T(t)\}$ on H such that $\|T(t)\| \leq Me^{\gamma t}$;

(b) $\omega \in L^2(0, T; H_n)$, where H_n is the Hilbert space where the noise takes values;

(c) $B: H_n \rightarrow H$, $C: H \rightarrow H_0$ (where H_0 is the observation Hilbert space) and $G: H_n \rightarrow H_0$ are bounded linear operators, with B H.S., such that $GB^* = 0$, $GG^* = I_{H_0}$.

The precise meaning of (1.1) is

$$x(t) = T(t)x_0 + \int_0^t T(t-s)B\omega(s) ds.$$

If $L^2(0, T; H_n)$ is equipped with the standard Gauss cylinder measure (which corresponds to model ω as a white-noise process) and x_0 is a Gaussian random variable with mean vector m_0 and nuclear covariance P_0 , the filtering problem consists of finding the best linear estimate $\hat{x}(t)$ of $x(t)$, given $\{y(s); 0 \leq s \leq t\}$. It evolves according to the equation

$$\dot{\hat{x}}(t) = A\hat{x}(t) + P(t)C^*(y(t) - C\hat{x}(t)), \quad \hat{x}(0) = m_0,$$

where $P(t)$ is the unique self-adjoint, nonnegative-definite, strongly continuous solution of the following Riccati equation:

$$(1.2) \quad P(t)x = T(t)P_0T^*(t)x + \int_0^t T(t-s)(BB^* - P(s)C^*CP(s))T^*(t-s)x ds,$$

$x \in H$. If (1.2) admits a steady-state solution P_∞ , namely,

$$(1.3) \quad P_\infty x = T(t)P_\infty T^*(t)x + \int_0^t T(t-s)(BB^* - P_\infty C^*CP_\infty)T^*(t-s)x ds,$$

then P_∞ is the solution of the following ARE:

$$(1.4) \quad AP_\infty + P_\infty A^* - P_\infty C^*CP_\infty + BB^* = 0.$$

By arguing as in [20], equation (1.4) can be derived from (1.3) only for $x \in D(A^*)$, but (1.4) is justified because $AP_\infty + P_\infty A^*$ admits the unique bounded extension $P_\infty C^*CP_\infty - BB^*$ to all of H .

The H.S. property of B has been assumed because it corresponds to the physically meaningful assumption of a spatially smoothed input noise. This hypothesis implies that the covariance operator of $B\omega(t)$ is the nuclear operator BB^* . On the other hand, also in the Ito formulation of the filtering problem, the distributed Wiener process is

defined with an incremental nuclear covariance operator. Regarding the observation noise, the white-noise approach has the advantage that the covariance operator is not required to be nuclear.

If the operator B is assumed to be nuclear, the ARE can be considered as an abstract equation in the Hilbert space of H.S. operators. Therefore, the smoothness property of its solution permits us to prove the convergence of the two approximating schemes here proposed toward the exact solution, under the unique assumption that the approximating semigroups converge in the Trotter–Kato sense. If the strong-smoothness action of H.S. operators [19, pp. 1289–1290] is exploited, no hypothesis on the finite-dimensional approximability of the adjoint semigroups is required. This makes such a class of approximation schemes much more feasible, because the choice of the approximating subspaces is crucial when the domain of the infinitesimal generator of the semigroup governing the system and the domain of its adjoint have a nondense intersection. This is the case, for instance, for hereditary systems [14], [17], [22], and [24]. In these cases the choice of approximating subspaces appears to be a nontrivial issue, because the simplest way of projecting the evolution equation on a finite-dimensional subspace does not give a convergent sequence for the corresponding solutions [14], [17], [22], and [24].

In this regard, it has recently been proved [5] that the spline scheme developed in [2] is such that the strong convergence property of the adjoint semigroups does not hold. A modified version of these splines was proposed in [24], where the optimal-control approximation problem was concerned. With reference to the same problem, the averaging approximation scheme was used in [22], where the trace-norm convergence of the solution of the resulting Riccati equation was obtained. However, in the above paper the approximation problem was dealt with on the basis of a conjecture; it assumed that the approximating systems are uniformly exponentially stable for sufficiently large dimensions if the original system is exponentially stable. This conjecture was shown to be correct for the averaging projection scheme in [35, §4.2].

The present paper is organized as follows. The discussion of the ARE in the Hilbert space of H.S. operators is given in §2. Existence and uniqueness of its solution are established together with a convergent sequence approximating the solution itself. Two kinds of finite-dimensional approximation schemes for the solution are given in §3. On the basis of recent results [19] concerning finite-dimensional approximability for the finite-horizon dynamical Riccati equation, the first method reduces the approximation problem to finding a large enough horizon time for approximating the steady-state solution. It constitutes an extension of the results in [19] and provides relaxed conditions for the approximability of the solution of the ARE. The second method computes the approximate solution by means of only algebraic linear operations, which can be easily implemented. It requires the assumption concerning the exponential stability of the perturbed semigroup. Section 4 contains a numerical example concerning the filtering problem for a delay system with approximating subspaces generated by first-order splines.

2. The algebraic Riccati equation in the space of Hilbert–Schmidt operators. Let H be a real, separable Hilbert space. An H.S. operator S on H is a bounded linear operator such that

$$\sum_{i=1}^{\infty} \|Se_i\|^2 := \|S\|_{\text{H.S.}}^2 < +\infty,$$

where $\{e_i\}$ is any orthonormal basis. The space $N(H)$ of all H.S. operators on H is a

Hilbert space with the inner product

$$[S, U]_{\text{H.S.}} = \sum_{i=1}^{\infty} (Se_i, Ue_i)$$

that is independent of the basis $\{e_i\}$ [1, p. 106]. The space of self-adjoint H.S. operators is a subspace of $N(H)$, which will be denoted by $N_s(H)$. The cone of nonnegative-definite operators in $N_s(H)$ will be denoted by $N_s^+(H)$. In the following, explicit reference to the space H will be omitted when this is clear from the context.

Let us recall that N is a (left and right) ideal [23, p. 148] of the space $L(H)$ of linear bounded operators on H such that

$$(2.1) \quad \|SL\|_{\text{H.S.}} \leq \|S\|_{\text{H.S.}} \|L\|, \quad S \in N, \quad L \in L(H),$$

and, furthermore,

$$(2.2) \quad \|S\|_{\text{H.S.}} = \|S^*\|_{\text{H.S.}}, \quad S \in N.$$

In this section the following ARE will be proved to have a unique solution in N_s^+ :

$$(2.3) \quad AP + PA^* - P\Sigma P + \Lambda = 0,$$

where

(H1) A is the infinitesimal generator of a strongly continuous semigroup of operators on H , $\{T(t), t \geq 0\}$, such that $\|T(t)\| \leq Me^{\lambda t}$;

(H2) Λ is trace-class operator in N_s^+ , and Σ is a bounded self-adjoint nonnegative-definite operator (not necessarily H.S.).

The hypotheses H1 and H2 will be referred to throughout the paper, and the next theorem provides a result that will be used later.

LEMMA 2.1. *Let $\{U(t), t \geq 0\}$ be a strongly continuous semigroup of operators on H , and let*

$$(2.4) \quad S(t)X = U(t)XU^*(t), \quad X \in N.$$

Then $\{S(t), t \geq 0\}$ is a strongly continuous semigroup on N .

Proof. Only the strong continuity needs to be shown. Let K be such that for h sufficiently small

$$(2.5) \quad \|U(h)\| \leq K,$$

and it follows by (2.1), (2.2), and (2.5) that

$$(2.6) \quad \begin{aligned} \|U(h)XU^*(h) - X\|_{\text{H.S.}}^2 &= \|U(h)X(U^*(h) - I) + (U(h) - I)X\|_{\text{H.S.}}^2 \\ &\leq 2K\|(U(h) - I)X^*\|_{\text{H.S.}}^2 + 2\|(U(h) - I)X\|_{\text{H.S.}}^2. \end{aligned}$$

Consider the first term of the last inequality in (2.6).

$$\begin{aligned} \|(U(h) - I)X^*\|_{\text{H.S.}}^2 &= \sum_{i=1}^{N_r} \|(U(h) - I)X^*\phi_i\|^2 + \sum_{i=N_r+1}^{\infty} \|(U(h) - I)X^*\phi_i\|^2 \\ &\leq \sum_{i=1}^{N_r} \|(U(h) - I)X^*\phi_i\|^2 + (1+K)^2 \sum_{i=N_r+1}^{\infty} \|X^*\phi_i\|^2 \\ &\leq \varepsilon, \end{aligned}$$

provided that N_r is chosen large enough that the second term is not greater than $\varepsilon/2$, whereas the first term can be made as small as needed by choosing h sufficiently small, because of the strong continuity of $U(h)$. The proof is completed by using the same arguments for the second term of (2.6). \square

LEMMA 2.2. *Let A be the infinitesimal generator of the strongly continuous uniformly bounded semigroup $\{U(t), t \geq 0\}$. Then the unique solution of the algebraic equation in N*

$$(2.7) \quad AX + XA^* + G = 0, \quad G \in N$$

admits a unique continuous extension on H given by

$$(2.8) \quad X = \int_0^\infty S(t)G \, dt,$$

where $S(t)$ is defined as in (2.4) provided that the integral (2.8) exists.

Proof. First of all, let us observe that $S(t)$ has the associate generator \mathcal{A} on N given by

$$(2.9) \quad \mathcal{A}X = AX + XA^*,$$

with

$$D(\mathcal{A}) \subseteq \{X \in N : R(X) \subset D(A)\},$$

which is dense because of the strong continuity of $\{S(t), t \geq 0\}$. Let us consider the equation in $C(0, \infty; N)$

$$(2.10) \quad \dot{X}(t) = \mathcal{A}X(t) + G, \quad X(0) \in N$$

that admits the mild solution [11, p. 41]

$$(2.11) \quad X(t) = S(t)X(0) + \int_0^t S(t-\tau)G \, d\tau, \quad t \geq 0.$$

Moreover,

$$X = \lim_{t \rightarrow \infty} X(t) = \int_0^\infty S(\tau)G \, d\tau$$

satisfies equation (2.7), as can be readily verified. \square

The existence theorem for the solution of equation (2.3) will be proved by constructing a sequence in N_s whose limit satisfies the ARE (2.3). For this we will use the following result concerning linear perturbed equations.

THEOREM 2.1. *Let $P \in N_s^+$, and let hypotheses $H1$ and $H2$ be satisfied. Moreover, let us assume that*

(i) *$(A - P\Sigma)$ is the infinitesimal generator of an asymptotically stable C_0 -semigroup $\{T_p(t), t \geq 0\}$ such that*

$$\int_0^\infty (T_p(t)(\Lambda + P\Sigma P)T_p^*(t)z, z) \, dt < \infty, \quad \forall z \in H.$$

(ii) *$(A, \Lambda^{1/2})$ is a controllable pair (in the sense specified in [10, p. 60]). Then the equation*

$$(2.12) \quad AX + XA^* - P\Sigma X - X\Sigma P + P\Sigma P + \Lambda = 0$$

admits a unique mild solution $X \in N_s^+$ such that

(a) *X is positive definite and*

(b) *$(A - X\Sigma)$ generates an asymptotically stable C_0 -semigroup $T_x(t)$ such that*

$$(2.13) \quad \int_0^\infty (T_x(t)(\Lambda + X\Sigma X)T_x^*(t)z, z) \, dt \leq (Xz, z).$$

Proof. Equation (2.12) can be rewritten as

$$(A - P\Sigma)X + X(A - P\Sigma)^* + P\Sigma P + \Lambda = 0,$$

which is of the kind (2.7), with $G = P\Sigma P + \Lambda \in N_S$. Hence from assumption (i) on the perturbed operator and by Lemma 2.1 it follows that the unique mild solution of equation (2.12) is

$$(2.14) \quad X = \int_0^\infty T_p(t)(\Lambda + P\Sigma P)T_p^*(t) dt, \quad X \in N_S.$$

The H.S. property of X is immediate. To prove the positive definiteness, let us assume the existence of a $z \in H$, $z \neq 0$, such that

$$\begin{aligned} 0 &= (Xz, z) = \int_0^\infty (T_p(t)(\Lambda + P\Sigma P)T_p^*(t)z, z) dt \\ &= \int_0^\infty \|\Lambda^{1/2}T_p^*(t)z\|^2 dt + \int_0^\infty \|\Sigma^{1/2}PT_p^*(t)z\|^2 dt, \end{aligned}$$

from which

$$(2.15) \quad \Lambda^{1/2}T_p^*(t)z = 0, \quad t \in [0, \infty],$$

$$(2.16) \quad \Sigma^{1/2}PT_p^*(t)z = 0, \quad t \in [0, \infty].$$

Moreover, by a well-known perturbation formula for semigroups [13, p. 69], $T_p^*(t)$ satisfies the following equation:

$$T_p^*(t)z = T^*(t)z - \int_0^t T^*(t-\tau)\Sigma PT_p^*(\tau)z d\tau.$$

Premultiplying both sides by $\Lambda^{1/2}$ and taking into account (2.15) and (2.16), we have

$$(2.17) \quad \Lambda^{1/2}T^*(t)z = 0,$$

against the hypothesis of approximate controllability [11, p. 60].

To prove assertion (b), let us consider the adjoint C_o -semigroup $\{T_x^*(t), t \geq 0\}$ generated by $(A - X\Sigma)^*$. Let $y(t) = T_x^*(t)z$, $z \in H$, and consider the following Lyapunov-like function:

$$(2.18) \quad V(y) = (Xy, y).$$

We have

$$\begin{aligned} \dot{V}(y) &= (X(A - X\Sigma)^*T_x^*(t)z, T_x^*(t)z) + (XT_x^*(t)z, (A - X\Sigma)^*T_x^*(t)z) \\ &= (T_x(t)(XA^* - X\Sigma X + AX - X\Sigma X)T_x^*(t)z, z), \end{aligned}$$

and by using (2.12) we obtain

$$\begin{aligned} \dot{V}(y) &= -(T_x(t)((P - X)\Sigma(P - X) + \Lambda + X\Sigma X)T_x^*(t)z, z) \\ (2.19) \quad &\leq -(\|\Lambda^{1/2}T_x^*(t)z\|^2 + \|\Sigma^{1/2}XT_x^*(t)z\|^2) \leq 0. \end{aligned}$$

To obtain inequality (2.13), let us write

$$0 \leq V(y(t)) = V(y(0)) + \int_0^t \dot{V}(y(\tau)) d\tau.$$

By (2.18) and (2.19) it follows that

$$0 \leq (Xz, z) - \int_0^t (T_x(\tau)(\Lambda + X\Sigma X)T_x^*(\tau)z, z) d\tau, \quad \forall t \geq 0,$$

which gives (2.13) by taking the limit for t going to infinity. \square

Now we can prove the uniqueness theorem for the steady-state solution of the ARE. For this purpose we need to state in advance the following lemma.

LEMMA 2.3. *Let $\{P_n\}$ be a sequence of nonnegative definite self-adjoint H.S. operators such that $P_n \leq P_0$. Then by denoting γ the maximum eigenvalue of P_0 we have*

$$P_n^2 \leq \gamma P_0.$$

Proof. It is enough to observe that for any $x \in H$

$$\begin{aligned} (P_n^2 x, x) &= (P_n P_n^{1/2} x, P_n^{1/2} x) \leq (P_0 P_n^{1/2} x, P_n^{1/2} x) \\ &\leq \gamma (P_n^{1/2} x, P_n^{1/2} x) = \gamma (P_n x, x) \\ &\leq \gamma (P_0 x, x). \end{aligned} \quad \square$$

THEOREM 2.2. *Assume that hypotheses H_1 and H_2 are verified. For a given $P_0 \in N_S^+$ let $(A - P_0 \Sigma)$ be the infinitesimal generator of a C_0 -semigroup $\{T_0(t), t \geq 0\}$ such that*

$$(2.20) \quad \int_0^\infty (T_0(t)(\Lambda + P_0 \Sigma P_0)T_0^*(t)z, z) dt = (P_1 z, z) < \infty, \quad \forall z \in H.$$

Then the sequence $\{P_n\}$ defined by

$$(2.21) \quad P_{n+1} = \int_0^\infty T_n(t)(\Lambda + P_n \Sigma P_n)T_n^*(t) dt, \quad n = 0, 1, \dots,$$

where $T_n(t)$ is generated by $(A - P_n \Sigma)$, converges in the H.S. norm toward the H.S. operator P_∞ , provided that $(A, \Lambda^{1/2})$ is a controllable pair. Moreover, if P_1 is a trace-class operator, so is P_∞ .

Proof. Equation (2.21) is well defined for $n = 0$. Moreover, by Theorem 2.1, $P_n, n \neq 0$, is such that

$$(P_n z, z) = \int_0^\infty (T_{n-1}(t)(\Lambda + P_{n-1} \Sigma P_{n-1})T_{n-1}^*(t)z, z) dt < \infty.$$

Hence (2.21) is well defined for each n , and the sequence $Q_n \in N_S^+$ given by

$$(2.22) \quad Q_n = P_n - P_{n+1}$$

can be defined. Now let us write (2.12) for $P = P_n$ and $X = P_{n+1}$ (respectively, $P = P_{n+1}$ and $X = P_{n+2}$). Then, let us subtract the second equation so obtained from the first. By defining $Q_{n+1} = P_{n+1} - P_{n+2}$, after simple calculations we get

$$AQ_{n+1} + Q_{n+1}A^* + Q_n \Sigma Q_n - P_{n+1} \Sigma Q_{n+1} - Q_{n+1} \Sigma P_{n+1} = 0,$$

which is again of the kind (2.12). Hence from (2.14)

$$(2.23) \quad Q_{n+1} = \int_0^\infty T_{n+1}(t)Q_n \Sigma Q_n T_{n+1}^*(t) dt,$$

where the integral exists, because from (2.18) and (2.19) we have

$$(Q_{n+1}z, z) \leq - \int_0^\infty \dot{V}(y) dt \leq (P_{n+1}z, z) < \infty, \quad \forall z \in H.$$

Moreover, (2.23) implies $Q_n \geq 0$, $n = 1, 2, \dots$, so that $P_n \geq P_{n+1}$. Let us define

$$(2.24) \quad P_\infty = P_0 - \sum_{i=0}^{\infty} Q_i.$$

Now it will be shown that $P_n \downarrow P_\infty$ in the H.S. norm. First one proves that $\|P_n\|_{\text{H.S.}} \downarrow \|P_\infty\|_{\text{H.S.}}$. Now let us recall that $A \geq B$, $A, B \in N_S$ implies that $\|A\|_{\text{H.S.}} \geq \|B\|_{\text{H.S.}}$, so we get

$$(2.25) \quad \|P_1\|_{\text{H.S.}} \geq \|P_n\|_{\text{H.S.}} \geq \|P_\infty\|_{\text{H.S.}}$$

and

$$(2.26) \quad \lim_{n \rightarrow \infty} \|P_n\|_{\text{H.S.}} = p \geq \|P_\infty\|_{\text{H.S.}}$$

On the other hand, we have

$$\begin{aligned} p^2 &= \lim_{n \rightarrow \infty} \sum_{i=1}^{\infty} (P_n \phi_i, P_n \phi_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^{\infty} (P_n^2 \phi_i, \phi_i) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^{N_\varepsilon} (P_n^2 \phi_i, \phi_i) + \lim_{n \rightarrow \infty} \sum_{i=N_\varepsilon+1}^{\infty} (P_n^2 \phi_i, \phi_i) \\ &\leq \lim_{n \rightarrow \infty} \sum_{i=1}^{N_\varepsilon} (P_n^2 \phi_i, \phi_i) + \sum_{i=N_\varepsilon+1}^{\infty} \gamma(P_0 \phi_i, \phi_i), \end{aligned}$$

where the last inequality follows from Lemma 2.3.

By the H.S. property of P_n we have that for a suitable choice of N_ε

$$\begin{aligned} p^2 &\leq \lim_{n \rightarrow \infty} \sum_{i=1}^{N_\varepsilon} (P_n^2 \phi_i, \phi_i) + \frac{\varepsilon}{2} = \sum_{i=1}^{N_\varepsilon} \lim_{n \rightarrow \infty} \|P_n \phi_i\|^2 + \frac{\varepsilon}{2} \\ &\leq \sum_{i=1}^{N_\varepsilon} \|P_\infty \phi_i\|^2 + \frac{\varepsilon}{2}, \end{aligned}$$

and therefore

$$(2.27) \quad p^2 < \|P_\infty\|_{\text{H.S.}}^2 + \varepsilon.$$

Since ε is arbitrary, from (2.26) and (2.27) it follows that

$$(2.28) \quad \lim_{n \rightarrow \infty} \|P_n\|_{\text{H.S.}} = p = \|P_\infty\|_{\text{H.S.}}$$

Now, from the parallelogram law, namely, $\|X - Y\|^2 + \|X + Y\|^2 = 2\|X\|^2 + 2\|Y\|^2$, by choosing $X = P_n$ and $Y = P_\infty$ we can write

$$\begin{aligned} (2.29) \quad \|P_n - P_\infty\|_{\text{H.S.}}^2 &= 2\|P_n\|_{\text{H.S.}}^2 + 2\|P_\infty\|_{\text{H.S.}}^2 - \|P_n + P_\infty\|_{\text{H.S.}}^2 \\ &= 2\|P_n\|_{\text{H.S.}}^2 - 2\|P_\infty\|_{\text{H.S.}}^2 + \|2P_\infty\|_{\text{H.S.}}^2 - \|P_n + P_\infty\|_{\text{H.S.}}^2. \end{aligned}$$

Moreover, $P_n + P_\infty \geq 2P_\infty$, so that

$$(2.30) \quad \|P_n + P_\infty\|_{\text{H.S.}} \geq \|2P_\infty\|_{\text{H.S.}}$$

Consequently, from (2.29) and (2.30) we have

$$\|P_n - P_\infty\|_{\text{H.S.}}^2 \leq 2(\|P_n\|_{\text{H.S.}}^2 - \|P_\infty\|_{\text{H.S.}}^2),$$

which by (2.28) implies that a large enough n can be found such that

$$\|P_n - P_\infty\|_{\text{H.S.}}^2 < \varepsilon. \quad \square$$

The nuclearity of P_∞ follows from the fact that $P_\infty \leq P_1$.

Remark 2.1. Note that Theorem 2.2 does not imply that $P_1 \leq P_0$, so if $P_0 = 0$ satisfies (2.20), it can be well assumed to start the iteration.

Remark 2.2. Inequality (2.20) is not equivalent to assuming that $T_0(t)$ is exponentially stable. This stronger property is guaranteed if it is also assumed that the pair $(A, \Lambda^{1/2})$ is exponentially stabilizable [1 p. 246]. In this case the nuclearity of P_1 is also guaranteed by the nuclearity of Λ . In the following we shall assume that P_1 is trace class.

To prove that P_∞ is the unique solution of the ARE, we need to state some preliminary results.

LEMMA 2.4 [11, Thm. 2.31]. *Let Ω be the infinitesimal generator of a C_0 -semigroup $\{\Gamma(t), t \geq 0\}$ on the Banach space B such that*

$$\|\Gamma(t)\| \leq Me^{\gamma t}, \quad \gamma \neq 0, t \geq 0,$$

and let P be a bounded linear operator on B . Then $(\Omega + P)$ is the infinitesimal generator of a C_0 -semigroup $\{\Gamma_p(t), t \geq 0\}$ such that

$$\|\Gamma_p(t)\| \leq Me^{(\gamma + \|P\|)t}, \quad t \geq 0.$$

LEMMA 2.5. *The following hold:*

$$(2.31) \quad \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|T_n(t) - T_\infty(t)\|_{\text{H.S.}}^2 = 0,$$

$$(2.32) \quad \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|T_n^*(t) - T_\infty^*(t)\|_{\text{H.S.}}^2 = 0,$$

where $T_n(t)$ is defined as in Theorem 2.2 and $T_\infty(t)$ is the semigroup generated by $A - P_\infty \Sigma$.

Proof. By Lemma (2.4) we have

$$(2.33) \quad \|T_n(t)\| \leq Me^{(\lambda + \|P_n \Sigma\|)t} \leq Me^{(\lambda + \|P_1\|_{\text{H.S.}} \|\Sigma\|)t} = Me^{\alpha t},$$

where $\alpha = \lambda + \|P_1\|_{\text{H.S.}} \|\Sigma\|$. The same bound (2.33) holds for $\|T_\infty(t)\|$. From the semigroup perturbation formula [13, p. 69] we have

$$T_\infty(t)x = T_n(t)x + \int_0^t T_n(t-\tau)(P_n - P_\infty)\Sigma T_\infty(\tau)x d\tau,$$

from which, after some calculations, it follows that

$$\|T_\infty(t) - T_n(t)\|_{\text{H.S.}}^2 \leq M^2 e^{2\alpha t} T^2 \|P_\infty - P_n\|_{\text{H.S.}}^2,$$

which proves (2.31) and (2.32) in light of Theorem 2.2 and equation (2.2). \square

LEMMA 2.6. *The operator P_∞ of Theorem 2.2 satisfies the following equation:*

$$(2.34) \quad P_\infty = \int_0^T T_\infty(t)(\Lambda + P_\infty \Sigma P_\infty) T_\infty^*(t) dt + T_\infty(T) P_\infty T_\infty^*(T), \quad T > 0.$$

Moreover, the operators in N_s^+ , $\mathcal{V} = \int_0^\infty T_\infty(t)(\Lambda + P_\infty \Sigma P_\infty) T_\infty^*(t) dt$, and P_∞ are positive definite.

Proof. From

$$P_{n+1} = \int_0^T T_n(t)(\Lambda + P_n \Sigma P_n) T_n^*(t) dt + \int_T^\infty T_n(t)(\Lambda + P_n \Sigma P_n) T_n^*(t) dt$$

we easily obtain

$$\begin{aligned} P_{n+1} &= \int_0^T T_n(t)(\Lambda + P_n \Sigma P_n) T_n^*(t) dt + T_n(T) \left\{ \int_0^\infty T_n(t)(\Lambda + P_n \Sigma P_n) T_n^*(t) dt \right\} T_n^*(T) \\ &= \int_0^T T_n(t)(\Lambda + P_n \Sigma P_n) T_n^*(t) dt + T_n(T) P_{n+1} T_n^*(T), \end{aligned}$$

which gives (2.34) by taking the limit for t going to infinity. From (2.34) it follows that \mathcal{V} is well defined because necessarily $\mathcal{V} \leq P_\infty \leq P_1$. By arguing as in (2.15)–(2.17) the positive definiteness of \mathcal{V} follows from the assumed approximate controllability of $(A, \Lambda^{1/2})$. This in turn implies the positive definiteness of P_∞ . \square

LEMMA 2.7. Let $Q(\tau) \in N_s^+$ be defined as

$$(2.35) \quad Q(\tau) = T_\infty(\tau)(\Lambda + P_\infty \Sigma P_\infty) T_\infty^*(\tau),$$

with P_∞ and $T_\infty(t)$ as in Theorem 2.2 and Lemma 2.5, respectively. Then \mathcal{V} is positive definite if and only if

$$(2.36) \quad \forall x \in H, \quad \exists \tau \geq 0: (Q(\tau)x, x) > 0.$$

Proof. If $\mathcal{V} > 0$, then $\forall x \in H$

$$0 < (\mathcal{V}x, x) = \int_0^\infty (Q(\tau)x, x) d\tau,$$

which implies (2.36). On the other hand, if (2.36) is satisfied for a $\bar{\tau} > 0$, then by the strong continuity of the semigroup there exists a $\sigma > 0$ such that $(Q(\bar{\tau} + \varepsilon)x, x) > 0$, $\forall \varepsilon \in [0, \sigma]$, which guarantees the positive definiteness of \mathcal{V} . \square

Since $Q(\tau) \in N_s^+$, $\forall \tau > 0$, $\forall x \in H$, it admits the representation

$$(2.37) \quad Q(\tau)x = \sum_{i=1}^\infty \lambda_i(\tau)(x, u_i^{(\tau)})u_i^{(\tau)},$$

with

$$\sum_i \lambda_i^2(\tau) < \infty, \quad \lambda_i(\tau) \geq 0 \quad i = 1, \dots,$$

where $\{u_i^{(\tau)}\}$ is a complete orthonormal eigenfunctions system.

Denoting the set of nonnegative rationals by Z^+ , we state the following lemma.

LEMMA 2.8. With the usual notations,

$$(2.38) \quad \text{span}\{u_i^r, i = 1, 2, \dots, r \in Z^+ : \lambda_i(r) > 0\} \equiv H$$

provided that the pair $(A, \Lambda^{1/2})$ is a controllable pair.

Proof. Let us suppose there exists a nonnull $h \in H$ such that

$$(2.39) \quad (h, u_i^r) = 0, \quad i = 1, 2, \dots \quad \forall r \in Z^+, \quad \lambda_i(r) > 0.$$

Hence by (2.38) and (2.39)

$$\begin{aligned}\sum_i (h, \sqrt{\lambda_i(r)} u_i^r)^2 &= \sum_i (h, Q^{1/2}(r) u_i^r)^2 \\ &= \|Q^{1/2}(r)h\|^2 = (Q(r)h, h) = 0.\end{aligned}$$

This implies $(Q(r)h, h) = 0 \forall r \in \mathbb{R}^+$ because Z^+ is dense in \mathbb{R}^+ , and hence $(\mathcal{V}h, h) = 0$, against the hypothesis of controllability by Lemma 2.6. \square

THEOREM 2.3. *Let $\{T_\infty(t), t \geq 0\}$ be the semigroup generated by $A_\infty = A - P_\infty \Sigma$. Then there exists an orthonormal basis $\{u_i\}$ on H such that*

$$(2.40) \quad \lim_{t \rightarrow \infty} \|T(t)u_i\| = 0.$$

Proof. First, let us prove that

$$(2.41) \quad \lim_{t \rightarrow \infty} \|T_\infty(t)u_i^r\| = 0 \quad \forall u_i^r: \lambda_i(r) > 0.$$

We have

$$\begin{aligned}(P_\infty x, x) &\geq \int_r^\infty (T_\infty(t-r)Q(r)T_\infty^*(t-r)x, x) dt \\ &= \int_r^\infty (Q(r)T_\infty^*(t-r)x, T_\infty^*(t-r)x) dt \\ &= \int_r^\infty \sum_i \lambda_i(r) (T_\infty^*(t-r)x, u_i^r)^2 dt.\end{aligned}$$

Now, for any orthonormal sequence $\{\phi_l\}$ on H we have

$$\begin{aligned}\text{tr } P_\infty &\geq \sum_l (P_\infty \phi_l, \phi_l) = \sum_l \int_r^\infty \sum_i \lambda_i(r) (T_\infty^*(t-r)\phi_l, u_i^r)^2 dt \\ &= \sum_i \lambda_i(r) \int_r^\infty \|T_\infty^*(t-r)u_i^r\|^2 dt.\end{aligned}$$

Since it must be that $\text{tr } P_\infty < \infty$, (2.41) is easily seen to hold.

Let us denote by $\{\tilde{u}_i, i = 1, 2, \dots\}$ a suitable renumbering of $\{u_i^r, i = 1, 2, \dots; r \in Z^+: \lambda_i(r) > 0\}$. Finally, by denoting by $\{u_l, l = 1, 2, \dots\}$ a basis on H obtained by the Gram-Schmidt orthogonalization procedure applied to $\{\tilde{u}_i, i = 1, 2, \dots\}$, equation (2.40) follows. \square

At this point the main theorem can be proved.

THEOREM 2.4. *Let $\{T_\infty(t), t \geq 0\}$ be defined as before, and suppose there exists $M_1 < \infty$ such that $\|T_\infty(t)\| \leq M_1$. Then for any operator $\Theta \in N_S^+$ the following equation holds:*

$$(2.42) \quad \lim_{t \rightarrow \infty} \|T_\infty(t)\Theta\|_{\text{H.S.}}^2 = 0.$$

Proof. By definition of H.S. norm one has

$$\|T_\infty(t)\Theta\|_{\text{H.S.}}^2 = \sum_i \|T_\infty(t)\Theta\phi_i\|^2 = \sum_i (T_\infty(t)\Theta\phi_i, T_\infty(t)\Theta\phi_i).$$

For $\{u_l\}$ defined as in Theorem 2.3 we have

$$\Theta\phi_i = \sum_l (\Theta\phi_i, u_l)u_l,$$

and substituting in the above expression we obtain

$$\begin{aligned}
 \|T_\infty(t)\Theta\|_{\text{H.S.}}^2 &= \sum_i \sum_l (\Theta\phi_i, u_l)(T_\infty(t)u_l, T_\infty(t)\Theta\phi_i) \\
 &= \sum_i \sum_{l=1}^{\bar{N}} (\Theta\phi_i, u_l)(T_\infty(t)u_l, T_\infty(t)\Theta\phi_i) \\
 &\quad + \sum_i \sum_{l=\bar{N}+1}^{\infty} (\Theta\phi_i, u_l)(T_\infty(t)u_l, T_\infty(t)\Theta\phi_i) \\
 &\leq \sum_{l=1}^{\bar{N}} \sum_i |(\Theta\phi_i, u_l)| \cdot |(\Theta T_\infty^*(t)T_\infty(t)u_l, \phi_i)| \\
 (2.43) \quad &\quad + \left(\sum_i \sum_{l=\bar{N}+1}^{\infty} (\Theta\phi_i, u_l)^2 \right)^{1/2} \left(\sum_i \sum_{l=\bar{N}+1}^{\infty} (\Theta T_\infty^*(t)T_\infty(t)u_l, \phi_i)^2 \right)^{1/2} \\
 &\leq \sum_{l=1}^{\bar{N}} \|\Theta u_l\| \cdot \|\Theta T_\infty^*(t)T_\infty(t)u_l\| \\
 &\quad + \left(\sum_{l=\bar{N}+1}^{\infty} \|\Theta u_l\|^2 \right)^{1/2} \left(\sum_{l=\bar{N}+1}^{\infty} \|\Theta T_\infty^*(t)T_\infty(t)u_l\|^2 \right)^{1/2} \\
 &\leq \|\Theta\|^2 \cdot M_1 \cdot \sum_{l=1}^{\bar{N}} \|T_\infty(t)u_l\| + \|\Theta\|_{\text{H.S.}} \cdot M_1^2 \cdot \left(\sum_{l=\bar{N}+1}^{\infty} \|\Theta u_l\|^2 \right)^{1/2}.
 \end{aligned}$$

Since Θ is an H.S. operator, we can choose \bar{N}_ε such that

$$\|\Theta\|_{\text{H.S.}} \cdot M_1^2 \cdot \left(\sum_{l=\bar{N}_\varepsilon+1}^{\infty} \|\Theta u_l\|^2 \right)^{1/2} < \frac{\varepsilon}{2},$$

and as a consequence of Theorem 2.3 we can find T_ε in order to obtain

$$\|\Theta\|^2 \cdot M_1 \cdot \sum_{l=1}^{\bar{N}_\varepsilon} \|T_\infty(t)u_l\| < \frac{\varepsilon}{2} \quad \forall t > T_\varepsilon,$$

so that from (2.43)

$$\|T_\infty(t)\Theta\|_{\text{H.S.}}^2 < \varepsilon \quad \forall t > T_\varepsilon,$$

which proves the theorem. \square

COROLLARY 2.1. *The semigroup $T_\infty(t)$, $t \geq 0$, is strongly stable.*

Proof. For each $z \in H$, $z \neq 0$, let Θ_z be the H.S. projection operator defined as $\Theta_z x = (x, z)z$, and let us choose an orthonormal basis $\{\phi_i\}$ on H such that $\phi_1 = z/\|z\|$. We have

$$\|T_\infty(t)\Theta_z\|_{\text{H.S.}}^2 = \sum_{i=0}^{\infty} \|T_\infty(t)\Theta_z\phi_i\|^2 = \|T_\infty(t)z\|^2,$$

which proves the strong stability of $T_\infty(t)$. \square

At this point the main theorem of this section can be stated.

THEOREM 2.5. *Let P_n be as in Theorem 2.2, and let P_1 be trace class. Then the trace-class operator P_∞ given by (2.34) is the unique solution of the ARE.*

$$AP_\infty + P_\infty A^* - P_\infty \Sigma P_\infty + \Lambda = 0.$$

Proof. By (2.34) we have

$$\left\| P_\infty - \int_0^T T_\infty(t)(\Lambda + P_\infty \Sigma P_\infty) T_\infty^*(t) dt \right\|_{\text{H.S.}} = \|T_\infty(\tau)P_\infty^{1/2}\|_{\text{H.S.}}^2,$$

where the right-hand side goes to zero for T going to infinity by Theorem 2.4 because $P_\infty^{1/2} \in N$. Hence

$$(2.44) \quad P_\infty = \int_0^\infty T_\infty(t)(\Lambda + P_\infty \Sigma P_\infty) T_\infty^*(t) dt.$$

Now, following the reasoning of [22, pp. 557–558], where the dual control problem is considered, one has that P_∞ maps the domain of A^* into the domain of A . Hence, by differentiating (2.44) we obtain

$$(2.45) \quad 0 = T_\infty(t)[AP_\infty + P_\infty A^* + \Lambda - P_\infty \Sigma P_\infty] T_\infty^*(t) \quad \forall t \geq 0.$$

Equation (2.45) implies that P_∞ is a solution of the ARE.

For the uniqueness let P and S be two such solutions, and let $\{T_q(t), t \geq 0\}$ and $\{T_p(t), t \geq 0\}$ be the semigroups generated by $(A - Q\Sigma)$ and $(A - P\Sigma)$, respectively. We find that the operator $E = P - Q$ satisfies

$$(A - Q\Sigma)E + E(A - Q\Sigma)^* - E\Sigma E = 0,$$

from which

$$(2.46) \quad E = - \int_0^\infty T_q(t) E \Sigma E T_q^*(t) dt.$$

Considering now the operator $F = Q - P$ and arguing as before, we readily obtain

$$(2.47) \quad F = - \int_0^\infty T_p(t) F \Sigma F T_p^*(t) dt.$$

From (2.46) and (2.47) it follows that $E = P - Q < 0$ and $F = Q - P < 0$. Hence it must be that $P = Q$. \square

3. Approximation methods.

3.1. Dynamic approximation. This approximation method is based on hypotheses H1 and H2 of § 2 and the trace-class property of the operator P_1 defined in Theorem 2.2. Moreover, it is assumed that $\forall x \in H$

$$(3.1) \quad \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|\Pi_n T(t)x - T_n(t)\Pi_n x\| = 0,$$

where

(i) Π_n is a projection operator on the finite-dimensional subspace $H_n \subset H$ such that $R(\Pi_n^*) \subset D(A)$ and $\Pi_n^* \Pi_n$ is strongly convergent to the identity on H . Therefore, $\|\Pi_n\| \leq M < \infty \forall n$ and for a suitably chosen M .

(ii) $T_n(t)$ is the semigroup generated by $\Pi_n A \Pi_n^*$.

Necessary and sufficient conditions for (3.1) to hold are established in the Trotter-Kato theorem (see, e.g., [31, p. 87]).

To state the convergence results for this approximation method we need to recall some results concerning the finite-horizon approximation of the Riccati equation.

Let $P(t)$ be the solution of the Riccati equation in $C(0, T; N_s^+(H))$,

$$(3.2) \quad P(t) = T(t)P(0)T^*(t) + \int_0^t T(t-s)[\Lambda - P(s)\Sigma P(s)]T(t-s)^* ds,$$

where Λ and Σ are as in hypothesis H2, $P(0)$ is trace class, and $P^{[n]}(t)$ is the solution

on $C(0, T; N_s^+(H_n))$ of

$$(3.3) \quad \begin{aligned} P^{[n]}(t) &= T_n(t) \Pi_n P(0) \Pi_n^* T_n^*(t) \\ &+ \int_0^t T_n(t-s) [\Pi_n A \Pi_n^* - P^{[n]}(s) \Pi_n \Sigma \Pi_n^* P^{[n]}(s)] T_n^*(t-s) ds. \end{aligned}$$

Then the following theorem holds [19]:

THEOREM 3.1. *Under hypothesis (3.1), for any $T > 0$*

$$(3.4) \quad \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|P^{[n]}(t) - \Pi_n P(t) \Pi_n^*\|_{\text{H.S.}} = 0.$$

This allows us to prove the following theorem:

THEOREM 3.2. *For each $\varepsilon > 0$ there exists a T_ε such that for any $T > T_\varepsilon$ an integer n_T exists such that for any $n > n_T$ the following holds:*

$$(3.5) \quad \|P^{[n]}(T) - \Pi_n P_\infty \Pi_n^*\|_{\text{H.S.}} < \varepsilon,$$

where $P^{[n]}(T)$ is the solution of (3.3) and P_∞ is the solution of the ARE.

Proof.

$$(3.6) \quad \|P^{[n]}(T) - \Pi_n P_\infty \Pi_n^*\|_{\text{H.S.}} \leq \|P^{[n]}(T) - \Pi_n P(T) \Pi_n^*\|_{\text{H.S.}} + M^2 \|P(T) - P_\infty\|_{\text{H.S.}}.$$

As in [19, p. 1299], the duality with the infinite-horizon control problem allows us to exploit the results stated in [22, Thm. 4.3] to prove that the following inequality holds:

$$(3.7) \quad 0 \leq P(t, \theta) - P_\infty \leq T_\infty(t) \theta T_\infty^*(t), \quad t > 0,$$

where $P(t, \theta)$ is the solution of (3.2) with $\theta = P(0)$, $\theta > P_\infty$ and is trace class, and $T_\infty(t)$ is the semigroup generated by $(A - P_\infty \Sigma)$. By (3.7) and Theorem 2.5 a sufficiently large T_ε can be found such that $\forall T > T_\varepsilon$

$$(3.8) \quad \|P(T, \theta) - P_\infty\|_{\text{H.S.}} < \varepsilon/2M^2.$$

By Theorem 3.1 it follows that $\forall T > T_\varepsilon$ there exists a n_T such that $\forall n > n_T$

$$\|P^{[n]}(T) - \Pi_n P(T) \Pi_n^*\|_{\text{H.S.}} < \varepsilon/2,$$

which, together with (3.8), implies (3.5). \square

Remark 3.1. Theorem 3.2 represents an extension of [19, Thm. 5] because of the quite restrictive hypothesis about the semigroup $T(t)$ generated by the operator A assumed there, i.e.,

$$\|T(t)\| \leq e^{-\omega t}, \quad \omega > \|A^{1/2}\| \|\Sigma^{1/2}\|$$

is now removed. In fact, any $\omega > 0$ is sufficient to guarantee that P_1 is trace class, so that Theorem 3.2 applies.

3.2. Algebraic approximation. This second approximation method is again based on hypotheses H1 and H2 of § 2, on the nuclearity of P_1 , and on the Trotter-Kato convergence of the approximating semigroup (3.1). Moreover, it is assumed that the semigroup $T_\infty(t)$ generated by $(A - P_\infty \Sigma)$ is exponentially stable, i.e.,

$$(3.9) \quad \|T_\infty(t)\| \leq e^{-\sigma t}, \quad \sigma > 0, \quad t \geq 0.$$

By the perturbation formula of semigroups (see Lemma 2.4) it can be shown that (3.9) implies that a constant λ exists such that

$$(3.10) \quad \|T(t)\| \leq e^{\lambda T}, \quad T \geq 0,$$

which is equivalent to the dissipativity of $A - \lambda I$. Inequality (3.10) implies that the semigroup $T_n(t)$ generated by $\Pi_n A \Pi_n^*$ is such that $\|T_n(t)\| \leq e^{\lambda t}$, $t \geq 0$, $n = 1, 2, \dots$ [37]. Let $\{P_n\}$ be the sequence of H.S. operators, decreasing to P_∞ , already computed in Theorem 2.2, and assume that P_0 is close enough to P_∞ to verify

$$\|P_0 - P_\infty\|_{\text{H.S.}} \cdot \|\Sigma\| < \sigma.$$

By the monotonicity of $\{P_n\}$ it follows that

$$(3.11) \quad \|P_n - P_\infty\|_{\text{H.S.}} \cdot \|\Sigma\| \leq \|P_0 - P_\infty\|_{\text{H.S.}} \cdot \|\Sigma\| < \sigma.$$

Note that such a choice of P_0 is always possible by assuming as P_0 the n th term of the sequence $\{P_n\}$ for a suitable n .

LEMMA 3.1. *The sequence of semigroups $\{T_k(t)$, $t \geq 0\}$ generated by the operators $(A - P_k)\Sigma$, $k = 0, 1, \dots$, satisfies the following inequality:*

$$(3.12) \quad \|T_k(t)\| \leq e^{-2\mu t}, \quad t \geq 0, \quad k = 0, 1, \dots,$$

where μ is a positive constant defined by

$$\mu = \frac{1}{2}(\sigma - \|(P_0 - P_\infty)\| \|\Sigma\|) > 0.$$

Proof. The proof is easily achieved with Lemma 2.4 by setting $\Omega = A - P_\infty \Sigma$ and $P = (P_\infty - P_k)\Sigma$ and by considering (3.11). \square

Now let $\{\Pi_n\}$ be a sequence of finite-dimensional orthoprojectors on H such that $\Pi_n H = H_n \subset D(A)$ strongly converges to the identity, and define

$$(3.13) \quad \Sigma_n = \Pi_n \Sigma \Pi_n, \quad A_n = \Pi_n A \Pi_n, \quad P_k^{(n)} = \Pi_n P_k \Pi_n.$$

Moreover, let Λ_n be such that

- (i) $\Lambda_n = \Lambda_n^T$;
- (ii) $\Lambda_n = \Pi_n \Lambda_n \Pi_n$;
- (iii) $\|\Lambda_n - \Lambda\|_{\text{H.S.}} \rightarrow 0$ as $n \rightarrow \infty$;
- (iv) $(A_n, \Lambda_n^{1/2})$ is a controllable couple $\forall n$.

A possible choice of Λ_n satisfying (i)–(iv) is

$$\Lambda_n = \Pi_n (\Lambda + E_n) \Pi_n,$$

where

$$E_n x = \frac{1}{n} \sum_{i=1}^{\infty} \frac{(x, \phi_i)}{2^i} \phi_i, \quad x \in H$$

for any orthonormal basis $\{\phi_i\}$ on H . In fact, E_n is a full-rank operator whose H.S. norm is $1/n$.

For each n let $\{\bar{P}_k^{(n)}\}$ be a sequence of positive-definite self-adjoint finite-rank operators converging to $\bar{P}^{(n)}$ defined by

$$(3.14) \quad \bar{P}_{k+1}^{(n)} = \int_0^\infty \bar{T}_k^{(n)}(t) [\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}] \bar{T}_k^{(n)*}(t) dt, \quad k = 0, 1, \dots,$$

$$\bar{P}_0^{(n)} = P_0^{(n)},$$

where $\{\bar{T}_k^{(n)}(t)$, $t \geq 0\}$ is the semigroup generated by the operator $(A_n - \bar{P}_k^{(n)} \Sigma_n)$. The positive definiteness of $\bar{P}_k^{(n)}$ follows from (iv) as in Lemma 2.6. By Theorem 2.2 it follows that if $\bar{P}_k^{(n)}$ is well defined, then the same is true for $\bar{P}_{k+1}^{(n)}$ and we have

$$(3.15) \quad \bar{P}_{k+1}^{(n)} \leq \bar{P}_k^{(n)}.$$

Let $\{T_k^{(n)}(t), t \geq 0\}$ and $\{\hat{T}_k^{(n)}(t), t \geq 0\}$ be the semigroups generated by $(A_n - P_k^{(n)}\Sigma_n)$ and $(A_n - \Pi_n P_k \Sigma \Pi_n)$, respectively. By considering (3.14) we see that

$$(3.16) \quad \bar{T}_0^{(n)}(t) = T_0^{(n)}(t).$$

Note that condition (3.12) implies [37] the following uniform-growth condition on $\hat{T}_k^{(n)}(t)$:

$$\|\hat{T}_k^{(n)}(t)\| \leq e^{-2\mu t}, \quad t \geq 0.$$

So by Lemma 2.4

$$(3.17) \quad \|T_k^{(n)}(t)\| \leq e^{(-2\mu t)} \cdot e^{(\|\Pi_n P_k \Sigma \Pi_n - P_k^{(n)}\Sigma_n\| t)}$$

provided that we set $\Omega = A_n - \Pi_n P_k \Sigma \Pi_n$ and $P = \Pi_n P_k \Sigma \Pi_n - P_k^{(n)}\Sigma_n$. Now we can prove the following lemmas.

LEMMA 3.2. *Let P_k be defined as in Theorem 2.2, and let $P_k^{(n)}$ and Σ_n be defined as in (3.13). Then for each $\varepsilon > 0$ a n_ε exists such that*

$$(3.18) \quad \|P_k^{(n)} - P_k\|_{\text{H.S.}} < \varepsilon \quad \forall n > n_\varepsilon$$

uniformly with respect to $k = 1, 2, \dots$,

$$(3.19) \quad \|P_k \Sigma (I - \Pi_n)\|_{\text{H.S.}} < \varepsilon \quad \forall n > n_\varepsilon$$

uniformly with respect to $n = 1, 2, \dots$,

$$(3.20) \quad \|\Pi_n P_k \Sigma \Pi_n - P_k^{(n)}\Sigma_n\|_{\text{H.S.}} < \varepsilon \quad \forall n > n_\varepsilon$$

uniformly with respect to $k = 1, 2, \dots$.

Proof. After choosing an orthonormal basis $\{\varphi_i\}$ on H such that $\{\varphi_i, i = 1, \dots, n\}$ is a basis on H_n , by simple calculations one obtains

$$(3.21) \quad \begin{aligned} \|\Pi_n P_k \Pi_n - P_k\|_{\text{H.S.}} &\leq 2\|P_k(\Pi_n - I)\|_{\text{H.S.}} \\ &= 2\left(\sum_{i=n+1}^{\infty} \|P_k \varphi_i\|^2\right)^{1/2} = 2\left(\sum_{i=n+1}^{\infty} (P_k^2 \varphi_i, \varphi_i)\right)^{1/2}. \end{aligned}$$

Taking into account Lemma 2.3 and that $P_k \leq P_1$ by Theorem 2.2, one has that the right-hand side of (3.21) is bounded by $2(\gamma \Sigma_{i=n+1}^{\infty} (P_1 \varphi_i, \varphi_i))^{1/2}$, where γ is the maximum eigenvalue of P_1 , so from the trace-class property of P_1 it follows that a n'_ε can be found such that (3.18) is satisfied.

Analogously, we have

$$\begin{aligned} \|P_k \Sigma (I - \Pi_n)\|_{\text{H.S.}} &= \left(\sum_{i=n+1}^{\infty} \|P_k \Sigma \varphi_i\|^2\right)^{1/2} \\ &= \left(\sum_{i=n+1}^{\infty} (P_k^2 \Sigma \varphi_i, \Sigma \varphi_i)\right)^{1/2} \leq \left(\gamma \sum_{i=n+1}^{\infty} (\Sigma^* P_1 \Sigma \varphi_i, \varphi_i)\right)^{1/2}, \end{aligned}$$

which is less than ε for $n > n''_\varepsilon$, because $\Sigma^* P_1 \Sigma$ is trace class.

For (3.20) it is enough to observe that

$$\|\Pi_n P_k \Sigma \Pi_n - P_k^{(n)}\Sigma_n\|_{\text{H.S.}} \leq \|P_k(I - \Pi_n)\|_{\text{H.S.}} \|\Sigma\|,$$

which is less than ε for $n > n'''_\varepsilon$ as in (3.21). By defining $n_\varepsilon = \max\{n'_\varepsilon, n''_\varepsilon, n'''_\varepsilon\}$ the proof is achieved. \square

LEMMA 3.3. Let $\Gamma(t)$ and $\Gamma^{(n)}(t)$ be two semigroups on H and H_n , respectively, such that $\|\Gamma(t)\| \leq e^{\alpha t}$, $t \geq 0$, and $\|\Gamma_n(t)\| \leq e^{\alpha t}$, $t \geq 0$, and satisfying the Trotter-Kato convergence theorem:

$$(3.22) \quad \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|\Pi_n \Gamma(t)x - \Gamma_n(t)\Pi_n x\| = 0 \quad \forall x \in H.$$

Moreover, let P_k be defined as in Theorem 2.2, and let L be a linear operator $\in N(H)$. Then

$$(3.23) \quad \sup_{t \in [0, T]} \|(\Pi_n \Gamma(t) - \Gamma(t)\Pi_n)P_k\|_{\text{H.S.}} < \varepsilon \quad \forall n > n_\varepsilon$$

uniformly with respect to $k = 1, 2, \dots$,

$$(3.24) \quad \sup_{t \in [0, T]} \|(\Pi_n \Gamma(t) - \Gamma_n(t)\Pi_n)L\|_{\text{H.S.}} < \varepsilon \quad \forall n > m(\varepsilon, L).$$

Proof. For each P_k , by definition of H.S. norm we have

$$(3.25) \quad \begin{aligned} \|(\Pi_n \Gamma(t) - \Gamma_n(t)\Pi_n)P_k\|_{\text{H.S.}}^2 &= \sum_{i=1}^{N_{\varepsilon, k}} \|(\Pi_n \Gamma(t) - \Gamma_n(t)\Pi_n)P_k \varphi_i\|^2 \\ &+ \sum_{i=N_{\varepsilon, k}+1}^{\infty} \|(\Pi_n \Gamma(t) - \Gamma_n(t)\Pi_n)P_k \varphi_i\|^2, \end{aligned}$$

where φ_i is an orthonormal sequence on H . Let us choose $N_{\varepsilon, k}$ such that

$$\sum_{i=N_{\varepsilon, k}+1}^{\infty} \|P_k \varphi_i\|^2 < \frac{\varepsilon^2}{8} e^{-2\alpha t}.$$

So the second term on the right-hand side of (3.25) is less than $\varepsilon^2/4$. By (3.22) the finite summation in (3.25) can be made less than $\varepsilon^2/4$ by choosing n greater than a suitable integer $m(\varepsilon, P_k)$. Hence for each P_k we have

$$(3.26) \quad \sup_{t \in [0, T]} \|(\Pi_n \Gamma(t) - \Gamma^{(n)}(t)\Pi_n)P_k\|_{\text{H.S.}} \leq \varepsilon$$

for $n > m(\varepsilon, P_k)$.

Now, let K_ε be such that for any $K > K_\varepsilon$, $\|P_k - P_\infty\| < (\varepsilon/4)e^{-\alpha T}$. Then, for $K > K_\varepsilon$

$$\begin{aligned} &\|(\Pi_n \Gamma(t) - \Gamma^{(n)}(t)\Pi_n)P_k\|_{\text{H.S.}} \\ &\leq \|(\Pi_n \Gamma(t) - \Gamma^{(n)}(t)\Pi_n)(P_k - P_\infty)\|_{\text{H.S.}} + \|(\Pi_n \Gamma(t) - \Gamma^{(n)}(t)\Pi_n)P_\infty\|_{\text{H.S.}} \\ &\leq 2e^{\alpha T} \|P_k - P_\infty\|_{\text{H.S.}} + \|(\Pi_n \Gamma(t) - \Gamma^{(n)}(t)\Pi_n)P_\infty\|_{\text{H.S.}} \\ &< \varepsilon \quad \forall n > m(\varepsilon/2, P_\infty). \end{aligned}$$

By choosing $n_\varepsilon = \max\{m(\varepsilon, P_1), \dots, m(\varepsilon, P_{k_\varepsilon}), m(\varepsilon/2, P_\infty)\}$, (3.23) is achieved. Inequality (3.24) is a straightforward consequence of (3.26). \square

LEMMA 3.4. Let $\{T_k^{(n)}(t), t \geq 0\}$ and $\{T_k(t), t \geq 0\}$ be defined as before. Then an n_0 exists such that

$$(3.27) \quad \|T_k^{(n)}(t)\| \leq e^{-\mu t} \quad \forall n > n_0, \quad \forall k.$$

Moreover,

$$(3.28) \quad \sup_{t \in [0, T]} \|\Pi_n T_k(t)x - T_k^{(n)}(t)\Pi_n x\| \leq \varepsilon \quad \forall n > n_1$$

uniformly with respect to $k = 1, 2, \dots$.

Proof. Inequality (3.27) follows from (3.17) if we take into account (3.20) and choose $\varepsilon \leq \mu$. As far as (3.28) is concerned, by the perturbation formula we have $\forall x \in H$

$$\begin{aligned} T_k^{(n)}(t)\Pi_n x &= T^{(n)}(t)\Pi_n x - \int_0^t T^{(n)}(t-\tau)(\Pi_n P_k \Sigma \Pi_n) T_k^{(n)}(\tau)\Pi_n x \, d\tau, \\ T_k(t)x &= T(t)x - \int_0^t T(t-\tau)P_k \Sigma T_k(\tau)x \, d\tau, \end{aligned}$$

from which

$$\begin{aligned} & \|(\Pi_n T_k(t) - T_k^{(n)}(t)\Pi_n)x\| \\ & \leq \|(\Pi_n T(t) - T^{(n)}(t)\Pi_n)x\| \\ & \quad + \int_0^t \|(\Pi_n T(t-\tau)P_k \Sigma T_k(\tau) - T^{(n)}(t-\tau)(\Pi_n P_k \Sigma \Pi_n) T_k^{(n)}(\tau)\Pi_n)x\| \, d\tau \\ (3.29) \quad & = \|(\Pi_n T(t) - T^{(n)}(t)\Pi_n)x\| \\ & \quad + \int_0^t \|(\Pi_n T(t-\tau)P_k \Sigma - T^{(n)}(t-\tau)(\Pi_n P_k \Sigma \Pi_n)) T_k(\tau)x\| \, d\tau \\ & \quad + \int_0^t \|T^{(n)}(t-\tau)(\Pi_n P_k \Sigma \Pi_n)(\Pi_n T_k(\tau) - T_k^{(n)}(\tau)\Pi_n)x\| \, d\tau. \end{aligned}$$

The first term on the right-hand side of (3.29) is less than $\varepsilon_1 \, \forall n > n_{\varepsilon_1}$, by (3.1). For the second term we have

$$\begin{aligned} & \int_0^t \|(\Pi_n T(t-\tau)P_k \Sigma - T^{(n)}(t-\tau)(\Pi_n P_k \Sigma \Pi_n)) T_k(\tau)x\| \, d\tau \\ & \leq \int_0^t \|\Pi_n T(t-\tau)P_k - T^{(n)}(t-\tau)\Pi_n P_k\|_{\text{H.S.}} \cdot \|\Sigma \Pi_n\| \|T_k(\tau)x\| \, d\tau \\ & \quad + \int_0^t \|\Pi_n T(t-\tau)P_k \Sigma (I - \Pi_n)\|_{\text{H.S.}} \|T_k(\tau)x\| \, d\tau \\ & \leq \|\Sigma\| \cdot T \sup_{t \in [0, T]} \|(\Pi_n T(t) - T^{(n)}(t)\Pi_n)P_k\|_{\text{H.S.}} \|x\| \\ & \quad + T e^{\lambda T} \|x\| \|P_k \Sigma (I - \Pi_n)\|_{\text{H.S.}} < \varepsilon_1 \quad \forall n > n'_{\varepsilon_1} \end{aligned}$$

uniformly with respect to $k = 1, 2, \dots$, by (3.19) and (3.23) when we take into account that $T(t)$ and $T_n(t)$ satisfy the condition of Lemma 3.3.

For the third term of right-hand side of (3.29) we have, taking into account that $\|P_k\|_{\text{H.S.}} \leq \|P_0\|_{\text{H.S.}}$,

$$\begin{aligned} & \int_0^t \|T^{(n)}(t-\tau)(\Pi_n P_k \Sigma \Pi_n)(\Pi_n T_k(\tau) - T_k^{(n)}(\tau)\Pi_n)x\| \, d\tau \\ & \leq e^{\lambda T} \|P_0\|_{\text{H.S.}} \|\Sigma\| \int_0^T \|(\Pi_n T_k(\tau) - T_k^{(n)}(\tau)\Pi_n)x\| \, d\tau. \end{aligned}$$

So it follows that $\forall n > \max\{n_{\varepsilon_1}, n'_{\varepsilon_1}\}$

$$\begin{aligned} & \|(\Pi_n T_k(t) - T_k^{(n)}(t)\Pi_n)x\| \leq 2\bar{\varepsilon} + e^{\lambda T} \|P_0\|_{\text{H.S.}} \cdot \|\Sigma\| \\ & \quad \cdot \int_0^T \|(\Pi_n T_k(\tau) - T_k^{(n)}(\tau)\Pi_n)x\| \, d\tau. \end{aligned}$$

Finally, (3.28) follows from the Gronwall inequality if we choose ε_1 such that

$$\varepsilon = 2 \exp\{e^{\lambda T} \|P_0\|_{\text{H.S.}} \|\Sigma\| T\} \cdot \varepsilon_1. \quad \square$$

Before proving the final convergence results, we observe that from $\|P_k\|_{\text{H.S.}} \leq \|P_0\|_{\text{H.S.}}$ and by definition of $P_k^{(n)}$ and Σ_n it follows that inequality (3.17) implies

$$(3.30) \quad \|T_k^{(n)}(t)\| \leq e^{\alpha t}, \quad t \geq 0, \quad \alpha = 2(-\mu + \|P_0\|_{\text{H.S.}} \|\Sigma\|),$$

so by (3.12) it is also true that

$$(3.31) \quad \|T_k(t)\| \leq e^{\alpha t}, \quad t \geq 0.$$

Moreover, if we set $\Omega = A_n - P_k^{(n)} \Sigma_n$ and $P = (P_k^{(n)} - \bar{P}_k^{(n)}) \Sigma_n$, Lemma 2.4 and inequality (3.17) give

$$(3.32) \quad \bar{T}_k^{(n)}(t) \leq e^{(-2\mu + \|\Pi_n P_k \Sigma \Pi_n - P_k^{(n)} \Sigma_n\| + \|P_k^{(n)} - \bar{P}_k^{(n)}\| \Sigma_n) t},$$

which, after we take into account that $\bar{P}_k^{(n)} \leq \bar{P}_0^{(n)} = P_0^{(n)}$, implies

$$(3.33) \quad \bar{T}_k^{(n)}(t) \leq e^{-\beta t}, \quad t \geq 0, \quad \beta = -2(\mu + 2\|P_0\| \cdot \|\Sigma\|),$$

LEMMA 3.5. Assume that for a given k

$$(3.34) \quad \lim_{n \rightarrow \infty} \|\bar{P}_k^{(n)} - P_k^{(n)}\|_{\text{H.S.}} = 0.$$

Then

$$(3.35) \quad \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|\bar{T}_k^{(n)}(t) - T_k^{(n)}(t)\| = 0.$$

Proof. Again, using the perturbation formula yields

$$\begin{aligned} \|\bar{T}_k^{(n)}(t) - T_k^{(n)}(t)\| &= \left\| \int_0^t T_k^{(n)}(t-\tau) [(\bar{P}_k^{(n)} - P_k^{(n)}) \Sigma_n] \bar{T}_k^{(n)}(\tau) d\tau \right\| \\ &\leq \int_0^t \|T_k^{(n)}(t-\tau)\| \cdot \|\bar{T}_k^{(n)}(\tau)\| \cdot \|(\bar{P}_k^{(n)} - P_k^{(n)}) \Sigma_n\| d\tau. \end{aligned}$$

Moreover, from (3.34), for n large enough

$$\|(\bar{P}_k^{(n)} - P_k^{(n)}) \Sigma_n\| \leq \hat{\varepsilon} \quad \forall n > n_{\varepsilon, k}.$$

By (3.33) there exists a constant $M > 0$ such that

$$\sup_{t \in [0, T]} \|\bar{T}_k^{(n)}(t)\| = M.$$

Moreover, by (3.27) and (3.34) it follows that

$$\sup_{t \in [0, T]} \|\bar{T}_k^{(n)}(t) - T_k^{(n)}(t)\| \leq \sup_{t \in [0, T]} \hat{\varepsilon} M \int_0^t e^{-\mu(t-\tau)} d\tau < \frac{\hat{\varepsilon} M}{\mu},$$

which proves (3.35). \square

THEOREM 3.3. Let $P_k^{(n)}$ and $\bar{P}_k^{(n)}$ be defined by (3.13) and (3.14), respectively. Then for any k

$$(3.36) \quad \lim_{n \rightarrow \infty} \|P_k^{(n)} - \bar{P}_k^{(n)}\|_{\text{H.S.}} = 0 \quad \forall k.$$

Proof. From (3.14) and (3.16) we have

$$\begin{aligned}
 & \|\bar{P}_1^{(n)} - P_1^{(n)}\|_{\text{H.S.}} \\
 &= \left\| \int_0^\infty T_0^{(n)}(t)[\Lambda_n + P_0^{(n)}\Sigma_n P_0^{(n)}]T_0^{(n)*}(t) dt \right. \\
 &\quad \left. - \Pi_n \int_0^\infty T_0(t)[\Lambda + P_0\Sigma P_0]T_0^*(t) dt \Pi_n \right\|_{\text{H.S.}} \\
 &\leq \left\| \int_0^{T_\varepsilon} (T_0^{(n)}(t)[\Lambda_n + P_0^{(n)}\Sigma_n P_0^{(n)}]T_0^{(n)*}(t) - \Pi_n T_0(t)[\Lambda + P_0\Sigma P_0]T_0^*(t)\Pi_n) dt \right\|_{\text{H.S.}} \\
 &\quad + \left\| \int_{T_\varepsilon}^\infty T_0^{(n)}(t)[\Lambda_n + P_0^{(n)}\Sigma_n P_0^{(n)}]T_0^{(n)*}(t) dt \right\|_{\text{H.S.}} \\
 &\quad + \left\| \int_{T_\varepsilon}^\infty \Pi_n T_0(t)[\Lambda + P_0\Sigma P_0]T_0^*(t)\Pi_n dt \right\|_{\text{H.S.}}.
 \end{aligned}$$

From (3.12) and (3.27) we may choose an $n > n_0$ such that the above expression is less than or equal to

$$\begin{aligned}
 (3.37) \quad & \left\| \int_0^{T_\varepsilon} (T_0^{(n)}(t)[\Lambda_n + P_0^{(n)}\Sigma_n P_0^{(n)}]T_0^{(n)*}(t) - \Pi_n T_0(t)[\Lambda + P_0\Sigma P_0]T_0^*(t)\Pi_n) dt \right\|_{\text{H.S.}} \\
 & + \int_{T_\varepsilon}^\infty e^{-2\mu t}(\|\Lambda_n\|_{\text{H.S.}} + \|P_0^{(n)}\|_{\text{H.S.}}^2\|\Sigma_n\|) dt + \int_{T_\varepsilon}^\infty e^{-4\mu t}(\|\Lambda\|_{\text{H.S.}} + \|P_0\|_{\text{H.S.}}^2\|\Sigma\|) dt.
 \end{aligned}$$

Let us choose T_ε such that

$$\begin{aligned}
 & \|P_1^{(n)} - \bar{P}_1^{(n)}\|_{\text{H.S.}} \leq \frac{\varepsilon}{2} + \left\| \int_0^{T_\varepsilon} (T_0^{(n)}(t)[\Lambda_n + P_0^{(n)}\Sigma_n P_0^{(n)}]T_0^{(n)*}(t) \right. \\
 &\quad \left. - \Pi_n T_0(t)[\Lambda + P_0\Sigma P_0]T_0^*(t)\Pi_n) dt \right\|_{\text{H.S.}} \\
 &\leq \frac{\varepsilon}{2} + \int_0^{T_\varepsilon} \|T_0^{(n)}\Pi_n L_n \Pi_n T_0^{(n)*}(t) \\
 &\quad - \Pi_n T_0(t)L T_0^*(t)\Pi_n\|_{\text{H.S.}} dt,
 \end{aligned}$$

where

$$L = \Lambda + P_0\Sigma P_0, \quad L_n = \Lambda_n + P_0^{(n)}\Sigma_n P_0^{(n)}.$$

Note that

$$(3.38) \quad \lim_{n \rightarrow \infty} \|L - L_n\|_{\text{H.S.}} = 0$$

by the H.S. property of P_0 and by definition of Λ_n and $P_0^{(n)}$.

Hence by simple calculation we have

$$\begin{aligned}
 & \|P_1^{(n)} - \bar{P}_1^{(n)}\|_{\text{H.S.}} \leq \frac{\varepsilon}{2} + \int_0^{T_\varepsilon} (\|T_0^{(n)}(t)\| \|T_0^{(n)*}(t)\| \|L_n - L\|_{\text{H.S.}} \\
 &\quad + \|(T_0^{(n)}(t)\Pi_n - \Pi_n T_0(t))L\|_{\text{H.S.}} \cdot \|T_0^{(n)*}(t)\| \\
 &\quad + \|T_0(t)\| \|L(\Pi_n T_0^{(n)*}(t) - T_0^*(t)\Pi_n)\|_{\text{H.S.}}) dt.
 \end{aligned}$$

Now let us observe that because of (3.28), (3.30), and (3.31) the semigroups $T_k(t)$ and $T_k^{(n)}(t)$ satisfy the condition of Lemma 3.3, so by (3.24), property (2.2) of H.S. operators, and (3.38) it follows that the integral on the right-hand side of the last inequality is less than $\varepsilon/2$ for $n > n_\varepsilon$. Hence, given $\varepsilon > 0$, there exists a $n_\varepsilon > n_1$ such that $\forall n > n_\varepsilon$

$$(3.39) \quad \|P_1^{(n)} - \bar{P}_1^{(n)}\|_{\text{H.S.}} < \varepsilon.$$

Now let us suppose that property (3.36) holds for a given k . We will show the same is true for $k+1$. We have

$$\begin{aligned} & \|\bar{P}_{k+1}^{(n)} - P_{k+1}^{(n)}\|_{\text{H.S.}} \\ &= \left\| \int_0^\infty \bar{T}_k^{(n)}(t) [\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}] \bar{T}_k^{(n)*}(t) dt \right. \\ & \quad \left. - \Pi_n \int_0^\infty T_k(t) [\Lambda + P_k \Sigma P_k] T_k^*(t) dt \Pi_n \right\|_{\text{H.S.}} \\ &\leq \left\| \int_0^{T_\varepsilon} (\bar{T}_k^{(n)}(t) [\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}] \bar{T}_k^{(n)*}(t) - \Pi_n T_k(t) [\Lambda + P_k \Sigma P_k] T_k^*(t) \Pi_n) dt \right\|_{\text{H.S.}} \\ & \quad + \left\| \int_{T_\varepsilon}^\infty \bar{T}_k^{(n)}(t) [\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}] \bar{T}_k^{(n)*}(t) dt \right\|_{\text{H.S.}} \\ & \quad + \left\| \int_{T_\varepsilon}^\infty \Pi_n T_k(t) [\Lambda + P_k \Sigma P_k] T_k^*(t) \Pi_n dt \right\|_{\text{H.S.}}. \end{aligned}$$

By (3.32), taking into account (3.20), the assumption on the k th step, and the uniform boundedness of Σ_n , for n large enough we obtain

$$\|\bar{T}_k^{(n)}(t)\| \leq e^{(-2\mu+2\varepsilon)t} = e^{-\mu't}, \quad 0 < \mu' = 2\mu - 2\varepsilon.$$

Now, by taking the norm in the above integral expression, it follows that

$$\begin{aligned} \|\bar{P}_{k+1}^{(n)} - P_{k+1}^{(n)}\|_{\text{H.S.}} &\leq \left\| \int_0^{T_\varepsilon} (\bar{T}_k^{(n)}(t) [\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}] \bar{T}_k^{(n)*}(t) \right. \\ & \quad \left. - \Pi_n T_k(t) [\Lambda + P_k \Sigma P_k] T_k^*(t) \Pi_n) dt \right\|_{\text{H.S.}} \\ & \quad + \int_{T_\varepsilon}^\infty e^{-2\mu't} \|\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}\|_{\text{H.S.}} dt \\ & \quad + \int_{T_\varepsilon}^\infty e^{-4\mu't} (\|\Lambda\|_{\text{H.S.}} + \|P_0\|_{\text{H.S.}}^2 \|\Sigma\|) dt. \end{aligned}$$

Taking into account that $\bar{P}_k^{(n)} \leq \bar{P}_1^{(n)}$, we can find a suitable constant q such that $\forall n$ we have

$$\begin{aligned} \|\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}\|_{\text{H.S.}} &\leq \|\Lambda_n - \Lambda\|_{\text{H.S.}} + \|\Lambda\|_{\text{H.S.}} + \|\bar{P}_k^{(n)}\|_{\text{H.S.}}^2 \|\Sigma_n\| \\ &\leq \|\Lambda_n - \Lambda\|_{\text{H.S.}} + \|\Lambda\|_{\text{H.S.}} + (\|\bar{P}_1^{(n)} - P_1^{(n)}\|_{\text{H.S.}} + \|P_1^{(n)}\|_{\text{H.S.}})^2 \|\Sigma\| \leq q \end{aligned}$$

because Λ_n is uniformly converging to Λ in the H.S. norm and by (3.39). Hence a T_ε exists such that

$$(3.40) \quad \|\bar{P}_{k+1}^{(n)} - P_{k+1}^{(n)}\|_{\text{H.S.}} \leq \frac{\varepsilon}{2} + \left\| \int_0^{T_\varepsilon} (\bar{T}_k^{(n)}(t)[\Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)}] \bar{T}_k^{(n)*}(t) - \Pi_n T_k(t)[\Lambda + P_k \Sigma P_k] T_k^*(t) \Pi_n) dt \right\|_{\text{H.S.}}.$$

By defining

$$L_{n,k} = \Lambda_n + \bar{P}_k^{(n)} \Sigma_n \bar{P}_k^{(n)},$$

$$L_k = \Lambda + P_k \Sigma P_k,$$

we have

$$(3.41) \quad \lim_{n \rightarrow \infty} \|L_{n,k} - L_k\|_{\text{H.S.}} = 0.$$

To prove (3.41) it is enough to observe that

$$\|L_{n,k} - L_k\|_{\text{H.S.}} \leq \|\Lambda - \Lambda_n\|_{\text{H.S.}} + \|\bar{P}_k^{(n)} - P_k^{(n)}\|_{\text{H.S.}} + \|P_k^{(n)} - P_k\|_{\text{H.S.}}$$

and to take into account the convergence of Λ_n to Λ in the H.S. norm, the assumption on the k th step, and (3.18). Hence by arguing as in the derivation of (3.39) we have that

$$(3.42) \quad \begin{aligned} \|\bar{P}_{k+1}^{(n)} - P_{k+1}^{(n)}\|_{\text{H.S.}} &\leq \frac{\varepsilon}{2} + \int_0^{T_\varepsilon} (\|\bar{T}_k^{(n)}(t)\| \|\bar{T}_k^{(n)*}(t)\| \|L_{n,k} - L_k\|_{\text{H.S.}} \\ &\quad + \|(\bar{T}_k^{(n)}(t) \Pi_n - \Pi_n T_k(t)) L_k\|_{\text{H.S.}} \cdot \|\bar{T}_k^{(n)*}(t)\| \\ &\quad + \|T_k(t)\| \|L_k(\Pi_n \bar{T}_k^{(n)*}(t) - T_k^*(t) \Pi_n)\|_{\text{H.S.}}) dt. \end{aligned}$$

Moreover,

$$(3.43) \quad \begin{aligned} \|(\bar{T}_k^{(n)}(t) \Pi_n - \Pi_n T_k(t)) L_k\|_{\text{H.S.}} &\leq \|(\bar{T}_k^{(n)}(t) \Pi_n - T_k^{(n)}(t) \Pi_n) L_k\|_{\text{H.S.}} \\ &\quad + \|(T_k^{(n)}(t) \Pi_n - \Pi_n T_k(t)) L_k\|_{\text{H.S.}}. \end{aligned}$$

By (3.30), (3.33), and (3.35) the semigroups $\bar{T}_k^{(n)}(t)$ and $T_k^{(n)}(t)$ satisfy the condition of Lemma 3.3. Because the same is true for $T_k^{(n)}(t)$ and $T_k(t)$, it follows by (3.24) that the right-hand side of (3.43) is less than $\varepsilon/2 T_\varepsilon$ for $n > m(\varepsilon, L_k)$. So by taking into account property (2.2) of H.S. operators and (3.41), an $n_{\varepsilon,k}$ can be found such that the integral of (3.42) is less than $\varepsilon/2$ for $n > n_{\varepsilon,k}$.

So far we have shown that if property (3.36) holds at the k th step, it still holds at the $(k+1)$ th step. Consequently, by (3.39) the theorem is proved by induction. \square

THEOREM 3.4. *Let P_∞ be the solution of the ARE, and let $\bar{P}_k^{(n)}$ be defined by (3.14). Then, given $\varepsilon > 0$, we can choose k and n in order to obtain*

$$\|P_\infty - \bar{P}_k^{(n)}\| \leq \varepsilon.$$

Proof. By noting that

$$\begin{aligned} \|P_\infty - \bar{P}_k^{(n)}\|_{\text{H.S.}} &= \|P_\infty - P_k + P_k - P_k^{(n)} + P_k^{(n)} - \bar{P}_k^{(n)}\|_{\text{H.S.}} \\ &\leq \|P_\infty - P_k\|_{\text{H.S.}} + \|P_k - P_k^{(n)}\|_{\text{H.S.}} + \|P_k^{(n)} - \bar{P}_k^{(n)}\|_{\text{H.S.}} \end{aligned}$$

we see that the proof follows immediately if we exploit Theorem 2.2 for the first term, property (3.18) for the second term, and Theorem 3.3 for the last term. \square

4. Numerical results. As an example of application of the approximation techniques described in § 3, let us consider the filtering problem for linear hereditary systems, defined by the following equations:

$$(4.1) \quad \dot{z}(t) = \sum_{i=0}^d A_i z(t-h_i) + \int_{-r}^0 A_{01}(s) z(t+s) ds + B_0 \omega_1(t),$$

$$t \in [0, T], \quad z(0) = z_0, \quad z(s) = z_1(s), \quad -r \leq s < 0,$$

$$(4.2) \quad y(t) = C_0 z(t) + G_0 \omega_2(t),$$

where $0 = h_0 < \dots < h_d = r$, $z(t) \in \mathbb{R}^n$, $A_i \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, $i = 1, \dots, d$, $A_{01} \in L^2(-r, 0; \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n))$, $B_0 \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^n)$, $C_0 \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^q)$, $G_0 \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^q)$, and ω_1 and ω_2 are assumed to be standard \mathbb{R}^p -valued and \mathbb{R}^r -valued independent white-noise processes, respectively.

We consider the Hilbert space $M^2 = \mathbb{R}^n \times L_2(-r, 0; \mathbb{R}^n)$ with inner product

$$[(v_0, v_1), (u_0, u_1)] = v_0^T u_0 + \int_{-r}^0 v_1^T(s) u_1(s) g(s) ds,$$

where $g(s)$ is a step weighting function such that

$$g(s) = i \quad \text{for } -h_{d-i+1} \leq s < -h_{d-i}, \quad i = 1, \dots, d,$$

and we define the operator $A: \mathcal{D}(A) \rightarrow M^2$ by

$$\mathcal{D}(A) = \{z = (z_0, z_1) \in M^2 : z_1 \in H^1(-r, 0; \mathbb{R}^n), z_0 = z_1(0)\},$$

$$Az = \left(A_0 z_0 + \sum_{i=1}^d A_i z_1(-h_i) + \int_{-r}^0 A_{01}(s) z_1(s) ds, z_1 \right).$$

Then the operator $(A - \lambda I)$ is dissipative in M^2 , i.e., $[Az, z]_{M^2} \leq \lambda \|z\|_{M^2}^2$, $z \in \mathcal{D}(A)$, and generates a strongly continuous semigroup $T(t)$ on M^2 such that $\|T(t)z\|_{M^2} \leq e^{\lambda t} \|z\|_{M^2}$ [2]. Moreover, if $T^{(n)}(t)$ is an approximating semigroup, a constant $\bar{\gamma}$ exists such that $\|T^{(n)}(t)\| \leq \exp\{\bar{\gamma}t\}$.

By setting

$$B = \begin{bmatrix} B_0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \omega^T = [\omega_1 \quad \omega_2]^T \in L_2(0, T; \mathbb{R}^{p+r}), \quad x^T = (z_0, z_1)^T,$$

$$C = [C_0 \quad 0], \quad G = [0 \quad G_0],$$

equations (4.1) and (4.2) can be rewritten as

$$(4.3) \quad x(t) = T(t)x_0 + \int_0^t T(t-s)B\omega(s) ds, \quad t \in [0, T],$$

$$(4.4) \quad y(t) = Cx(t) + G\omega(t),$$

and the solution of (4.1) is given by the first component of (4.3). Without any loss of generality we can always assume that $G_0 G_0^T = I$; moreover, B is H.S. because of the finite-dimensionality of its range.

Now, given the mean vector and the covariance operator of the initial state x_0 , the best linear estimate of $x(t)$, and consequently of $z(t)$, can be obtained by the infinite-dimensional Kalman filter.

For each n let Π_n be the projection operator on the subspace V_n generated by the following piecewise linear splines:

$$\begin{aligned} v_{i,0}^{(n)}(t) &= 1_{[-r/n,0]}(t)\{1 + (nt/r)\}e_i, \\ v_{i,m}^{(n)}(t) &= -1_{[-mr/n, -(m-1)r/n]}(t)\{m-1 + (nt)/r\} \\ &\quad \cdot 1_{[-(m+1)r/n, -mr/n]}(t)\{m+1 + (nt)/r\}e_i, \quad m = 1, \dots, 2^n - 1, \\ v_{i,n}^{(t)} &= -1_{[-1, -1+r/n]}(t)\{n-1 + (nt)/r\}e_i, \quad i = 1, 2, \dots, n, \end{aligned}$$

where e_i , $i = 1, 2, \dots, n$, is the canonical basis of \mathbb{R}^n , and we are considering $\mathcal{D}(A) \equiv H^1(-r, 0; \mathbb{R}^n)$. For such a scheme, the convergence of the approximate semigroup toward the actual one holds [2].

It is worth recalling that, as shown in [5], the above scheme does not satisfy the strong convergence property for the adjoint semigroups. Nevertheless, it can be used in the present filtering problem because both of the approximating methods proposed here do not require the fulfillment of such a property.

For the matrix representation, with respect to the selected basis, of all the operators to be used in the implementation of both the approximation methods, we refer to [2] and [19]. The numerical example proposed consists of the state estimation for the system described by the following equations:

$$(4.5) \quad \dot{x}(t) = a_0 x(t) + a_1 x(t-1) + bw(t),$$

$$(4.6) \quad y(t) = cx(t) + gv(t),$$

with initial conditions $x(s) = -50s + 50$, $-1 \leq s < 0$, and with $w(t)$ and $v(t)$ being independent standard white-noise processes.

To satisfy all the hypotheses that make the approximation theorems of § 3 hold, we imposed the stability condition $a_0 + |a_1| < 0$ [30] by choosing $a_0 = -3$, $a_1 = 2$. Moreover, we set $b = 3.5$ and $d = 2$. The system dynamic was simulated according to the following difference equations:

$$\begin{aligned} x((k+1)\Delta) &= e^{a_0\Delta}x(k\Delta) + \frac{\Delta}{2}[a_1x((k+1)\Delta - 1) + e^{a_0\Delta}a_1x(k\Delta - 1)] + w_k, \\ y(k\Delta) &= x(k\Delta) + v_k, \end{aligned}$$

where $\Delta = 0.025$, $k = 0, 1, \dots, 80$, and $\{w_k\}$ and $\{v_k\}$ are independent zero-mean Gaussian white sequences, with covariances $b^2 \int_0^\Delta \exp\{2a_0 t\} dt$ and d^2 , respectively, which were generated by using NAG FORTRAN subroutine G05DDF.

The filter was initialized with $\hat{x}_n(0) = y(0)$. The estimate $\hat{x}_n(t)$ was determined at each time instant by using the gain operator obtained off-line from the approximate solution $P_\infty^{(n)}$ of the ARE. It was computed by exploiting the two methods proposed in § 3, for $n = 3, 5, 7$. We will refer first to the method of obtaining $P_\infty^{(n)}$ through the dynamical Riccati equation evolving toward the steady state; the second method is for computing $P_\infty^{(n)}$ by algebraic linear operations.

The numerical values $\hat{x}_n(t)$, $t = k(2\Delta)$, $k = 1, 2, \dots, 40$, have been computed by integrating the approximate filter equation by means of the NAG FORTRAN subroutine D02EAF, which uses a variable-order, variable-step Gear method. The same subroutine has been also used to integrate the approximate dynamical Riccati equation and to compute the integral in equation (3.14) with null initial condition at each iteration. For both the approximation methods the achievement of numerical stability has been tested within the tolerance limit of 10^{-5} for 10 successive iterations.

For each experiment the filter performance was evaluated by computing the error statistics

$$\sigma_p^2 = \frac{1}{40} \sum_{k=1}^{40} [x(k\Delta) - \hat{x}(k\Delta)]^2$$

and by defining the signal-to-noise ratio improvement (SNRI) as

$$\text{SNRI} := 10 \log_{10}(\text{variance of observation noise} / \sigma_p^2).$$

For the first method we obtained SNRI = 1.606, 4.153, 4.382 for $n = 3, 5, 7$, respectively, and for the second method we obtained SNRI = 1.609, 4.150, 4.378 for the same respective values of n .

The whole numerical example was carried out on a VAX 780 computer. The simulation results are reported in Figs. 1, 2, and 3, corresponding to schemes of order $n = 3, 5$, and 7 , respectively. In each figure plots (a) and (b) represent the behavior of the filter built according to the first and second method, respectively. The good agreement between the two approximation methods is also evident from a comparison of the numerical values reported in Table 1.

5. Concluding remarks. The problem of approximating the infinite-dimensional algebraic Riccati equation, as an abstract equation in the Hilbert space of H.S. operators, has been considered. Two methods have been proposed. The first one is based on the results established in a previous paper, where, by exploiting the approximability of the corresponding dynamical Riccati equation and its time convergence toward the steady state, the problem was reduced to finding a large enough time horizon to approximate the steady-state solution. Here this approximation scheme has been shown to converge under more general conditions than those in the previous setting, relaxing, in particular, the strong hypothesis concerning the exponential stability of the unperturbed semigroup.

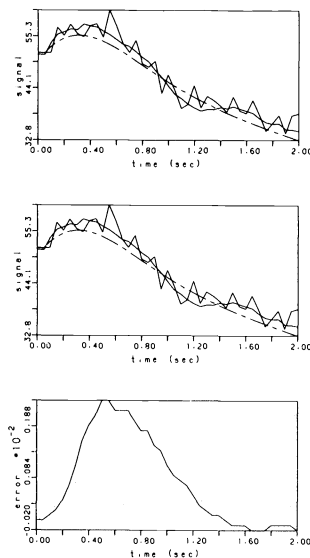


FIG. 1. Time plots of the noisy-state observation (continuous ragged line), true-state evolution (continuous smooth line), and approximate-Kalman-state estimation (dashed line) obtained with the (a) first and (b) second method with $n = 3$. (c) Time plot of the difference between the estimate values.

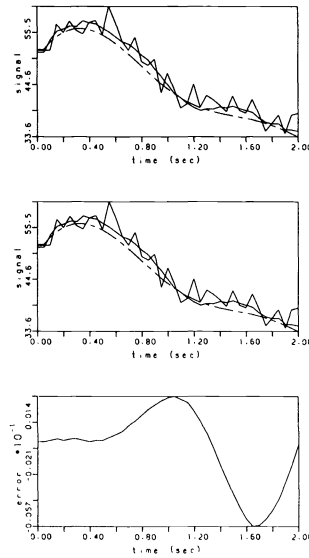


FIG. 2. Time plots of the noisy-state observation (continuous ragged line), true-state evolution (continuous smooth line), and approximate-Kalman-state estimation (dashed line) obtained with the (a) first and (b) second method with $n = 5$. (c) Time plot of difference between the estimate values.

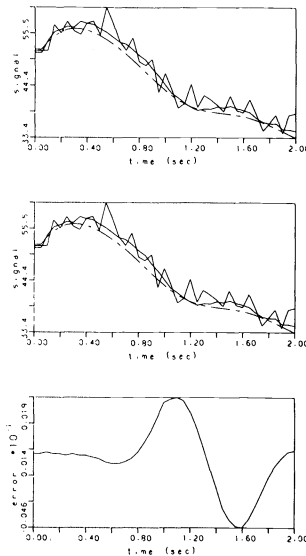


FIG. 3. Time plots of the noisy-state observation (continuous ragged line), true-state evolution (continuous smooth line), and approximate-Kalman-state estimation (dashed line) obtained with the (a) first and (b) second method with $n = 7$. (c) Time plot of the difference between the estimate values.

The second method provides a scheme for computing the approximate solution by means of only algebraic linear operations under the exponential-stability assumption for the perturbed semigroup. In fact, the solutions of a sequence of finite-dimensional linear equations have been proved to converge to the exact steady-state solution of the original problem.

TABLE 1

Approximate steady-state solution of the algebraic Riccati equation $P_\infty^{(n)}$ computed by using the first and second methods with schemes of order $n = 3, 5, 7$.

	1st Method	2nd Method
n	$P_\infty^{(n)}$	$P_\infty^{(n)}$
3	0.0204 0.0108 0.0092	0.0186 0.0088 0.0069
	0.0108 0.0151 0.0125	0.0088 0.0128 0.0098
	0.0092 0.0125 0.0138	0.0069 0.0098 0.0107
5	0.0215 0.0138 0.0092 0.0091 0.0100	0.0205 0.0127 0.0080 0.0079 0.0087
	0.0138 0.0195 0.0143 0.0087 0.0086	0.0127 0.0182 0.0129 0.0074 0.0071
	0.0092 0.0143 0.0196 0.0146 0.0096	0.0080 0.0129 0.0181 0.0131 0.0081
	0.0091 0.0087 0.0146 0.0182 0.0154	0.0079 0.0074 0.0131 0.0166 0.0137
	0.0100 0.0086 0.0096 0.0154 0.0168	0.0087 0.0071 0.0081 0.0137 0.0151
7	0.0217 0.0159 0.0117 0.0096 0.0084 0.0091 0.0102	0.0203 0.0144 0.0102 0.0081 0.0068 0.0074 0.0083
	0.0159 0.0202 0.0163 0.0117 0.0093 0.0082 0.0089	0.0144 0.0186 0.0148 0.0101 0.0075 0.0064 0.0069
	0.0117 0.0163 0.0198 0.0165 0.0120 0.0094 0.0087	0.0102 0.0148 0.0182 0.0149 0.0103 0.0076 0.0067
	0.0096 0.0117 0.0165 0.0195 0.0165 0.0115 0.0095	0.0081 0.0101 0.0149 0.0177 0.0147 0.0096 0.0074
	0.0084 0.0093 0.0120 0.0165 0.0200 0.0165 0.0119	0.0068 0.0075 0.0103 0.0147 0.180 0.0144 0.0097
	0.0091 0.0082 0.0094 0.0115 0.0165 0.0194 0.0170	0.0074 0.0064 0.0076 0.0096 0.0144 0.0171 0.0146
	0.0102 0.0089 0.0087 0.0095 0.0119 0.0170 0.0182	0.0083 0.0069 0.0067 0.0074 0.0097 0.0146 0.01157

Both the solving methods work under quite general conditions that, on the other hand, do not allow for a uniform convergence of the approximate solution toward the actual one.

Finally, we wish to point out that no assumption on the finite-dimensional approximability of the adjoint semigroup is made. Such a hypothesis, which is generally requested in the literature, is often difficult to verify when $D(A) \cap D(A^*)$ is not dense. This is the case, for instance, in hereditary systems.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1981.
- [2] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equation*, J. Differential Equations, 34 (1979), pp. 496–522.
- [3] H. T. BANKS AND C. WANG, *Optimal feedback control for infinite-dimensional parabolic evolution systems: approximation techniques*, SIAM J. Control Optim., 27 (1989), pp. 1182–1219.
- [4] H. T. BANKS, G. J. ROSEN, AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830–855.
- [5] J. BURNS, K. ITO, AND G. PROPST, *On nonconvergence of adjoint semigroups for control systems with delays*, SIAM J. Control Optim. 26 (1988), pp. 1441–1454.
- [6] A. BENSOUSSAN, *Filtrage optimal des systèmes linéaires*, Dunod, Paris, 1971.
- [7] R. CURTAIN, *Infinite-dimensional filtering*, SIAM J. Control Optim., 13 (1975), pp. 89–104.
- [8] R. F. CURTAIN, *The infinite-dimensional Riccati equation with applications to affine hereditary differential systems*, SIAM J. Control Optim., 13 (1975), pp. 1130–1143.
- [9] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation*, J. Math. Anal. Appl., 47 (1974), pp. 43–57.
- [10] ———, *The infinite-dimensional Riccati equation for systems defined by evolutions operators*, SIAM J. Control Optim., 14 (1976), pp. 951–983.
- [11] ———, *Infinite Dimensional Linear System Theory*, Springer-Verlag, Berlin, 1978.

- [12] G. DA PRATO, *Equations d'évolutions dans des algèbres d'opérateurs et application à des équations quasi-linéaires*, J. Math. Pures Appl., 48 (1969), pp. 59–107.
- [13] E. B. DAVIES, *One Parameter Semigroups*, Academic Press, London, 1980.
- [14] M. C. DELFOUR, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101–162.
- [15] ———, *The linear quadratic optimal control problem with delays in state and control variables: a state space approach*, SIAM J. Control Optim., 24 (1986), pp. 835–883.
- [16] M. C. DELFOUR AND R. S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, SIAM J. Control Optim., 10 (1972), pp. 298–328.
- [17] ———, *Hereditary differential systems with constant delays, I: General case*, J. Differential Equations, 12 (1972), pp. 213–235.
- [18] F. FLANDOLI, I. LASIECKA, AND R. TRIGGIANI, *Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler–Bernoulli boundary control problems*, Ann. Mat. Pura Appl., N153 (1988), pp. 307–382.
- [19] A. GERMANI, L. JETTO, AND M. PICCIONI, *Galerkin approximation for optimal filtering of infinite-dimensional linear systems*, SIAM J. Control Optim., 26 (1988), pp. 1287–1305.
- [20] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.
- [21] ———, *An analysis of optimal modal regulation: convergence and stability*, SIAM J. Control Optim., 19 (1981), pp. 686–707.
- [22] ———, *Linear quadratic optimal control of hereditary differential systems: infinite-dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 31 (1983), pp. 95–139.
- [23] W. GREUB, *Linear Algebra*, Springer-Verlag, Berlin, New York, 1975.
- [24] F. KAPPEL AND D. SALAMON, *Splines approximation for retarded Systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082–1117.
- [25] K. KUNISCH, *Approximation schemes for the linear-quadratic optimal control problem associated with delay equations*, SIAM J. Control Optim., 20 (1982), pp. 506–540.
- [26] I. LASIECKA AND R. TRIGGIANI, *Riccati equations for hyperbolic partial differential equations with $L_2(0, T; L_2(\Gamma))$ -Dirichlet boundary terms*, SIAM J. Control Optim., 24 (1986), pp. 884–925.
- [27] I. LASIECKA, *Approximation of Riccati equation for abstract boundary control problems—applications to hyperbolic systems*, Numer. Funct. Anal. Optim., 8 (1985/86), pp. 207–243.
- [28] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1971.
- [29] K. MITTER AND R. B. VINTER, *Filtering of linear stochastic hereditary differential systems*, in Numerical Methods and Computer-Systems Modelling, Lecture Notes in Economics and Mathematical Systems 107, Springer-Verlag, Berlin, New York, 1975.
- [30] T. MORI, *Criteria for asymptotic stability of linear time-delay systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 158–161.
- [31] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [32] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic control problem for infinite-dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.
- [33] I. G. ROSEN, *On Hilbert–Schmidt Norm Convergence of Galerkin Approximation for Operator Riccati Equations*, International Series of Numerical Mathematics 91, Birkhauser Verlag, Basel, Switzerland, 1989 pp. 335–349.
- [34] ———, *Convergence of Galerkin approximations for operator Riccati equations—A nonlinear evolution equation approach*, J. Math. Anal. Appl., 155 (1991), pp. 226–248.
- [35] D. SALAMON, *Structure and stability of finite dimensional approximations for functional differential equations*, SIAM J. Control Optim., 23 (1985), pp. 928–951.
- [36] R. B. VINTER, *Filter stability for stochastic evolution equations*, SIAM J. Control Optim., 15 (1977), pp. 465–485.
- [37] J. A. WALKER, *Dynamical Systems and Evolutions Equations*, Plenum Press, New York, 1980.
- [38] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, J. Appl. Math. Optim., 2 (1976), pp. 251–258.

INTERIOR DUAL LEAST 2-NORM ALGORITHM FOR LINEAR PROGRAMS*

RUDY SETIONO†

Abstract. An interior algorithm is proposed for solving the dual of the least 2-norm formulation of a linear program. This is a convex quadratic problem with nonnegativity constraints only. Sixty-six test problems, including 63 NETLIB problems, were solved very accurately. The total time speedup of the algorithm for all 66 problems over MINOS 5.3 is 2.10. Linear convergence of the algorithm is also established.

Key words. linear programming, interior point method, Newton step

AMS subject classifications. 90C05, 90C25

1. Introduction. It is well known [16], [18] that a linear program

$$(1) \quad \min_x cx \quad \text{s.t. } Ax \geq b, \quad x \geq 0$$

is solvable if and only if the quadratic program

$$(2) \quad \min_x cx + \frac{\varepsilon}{2} xx \quad \text{s.t. } Ax \geq b, \quad x \geq 0$$

is solvable by the same \bar{x} for all $\varepsilon \in (0, \bar{\varepsilon}]$ for some $\bar{\varepsilon} > 0$. If $x(\varepsilon)$ solves the quadratic problem (2), then it is the solution of the linear program (1) that is closest to the origin in the 2-norm. The dual of the above quadratic program [14] is

$$(3) \quad \max_{x, u, v} -\frac{\varepsilon}{2} xx + bu$$

subject to

$$(4) \quad \varepsilon x - A'u - v + c = 0,$$

$$(5) \quad u, v \geq 0.$$

Elimination of x from the dual problem by using the constraint relation

$$x = \frac{1}{\varepsilon} (A'u + v - c)$$

leads to the following exterior penalty function with penalty parameter ε associated with the dual of linear program (1):

$$(6) \quad \min_{u, v} \frac{1}{2} \|A'u + v - c\|^2 - \varepsilon bu \quad \text{s.t. } u, v \geq 0.$$

The Karush–Kuhn–Tucker optimality conditions for the quadratic problem (6) can be expressed as a symmetric linear complementarity problem

$$(7) \quad Mz + q \geq 0, \quad z \geq 0, \quad z(Mz + q) = 0$$

* Received by the editors August 27, 1990; accepted for publication (in revised form) February 20, 1992.

† Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, Wisconsin 53706. This research was supported by National Science Foundation grants DCR-8521228 and CCR-8723091 and U.S. Air Force Office of Scientific Research grants AFOSR-86-0172 and AFOSR-89-0410. Current address, Department of Information Systems and Computer Science, National University of Singapore, Singapore 0511.

after the following identifications have been made:

$$(8) \quad M := \begin{pmatrix} AA' & A \\ A' & I \end{pmatrix}, \quad q := \begin{pmatrix} -Ac - \varepsilon b \\ -c \end{pmatrix}, \quad z := \begin{pmatrix} u \\ v \end{pmatrix}.$$

Iterative successive overrelaxation methods have been proposed for solving the symmetric linear complementarity problems [15]. A successive overrelaxation method that preserves the sparsity structure of the problem has been implemented to solve very large linear programs [3], [19]. These large linear programs with up to 125,000 constraints and 500,000 variables are impossible to solve by using a direct method such as the simplex.

Our approach to finding the least 2-norm solution of a linear program is to use an interior penalty function. Since the only constraints present in the dual problem (6) are nonnegativity constraints, an initial starting point for the algorithm can be obtained trivially. The interiority of the iterates are easy to maintain by taking an appropriate step size. These facts constitute the motivation behind our dual interior penalty method.

We now briefly outline the contents of the paper. In § 2 we describe the algorithm, and in § 3 we establish its linear convergence. In § 4 we give computational results. In § 5 we summarize the paper.

2. Interior dual least 2-norm (IDLN) algorithm. We consider the linear program given in the standard form

$$(9) \quad \min_x cx \quad \text{s.t.} \quad Ax = b, \quad x \geq 0$$

and its dual

$$(10) \quad \max_{u,v} bu \quad \text{s.t.} \quad A'u + v = c, \quad v \geq 0,$$

where A is an $m \times n$ matrix; c , x , v are n -vectors; and b , u are m -vectors. We make the following assumption throughout regarding this linear program.

Assumption 1. The dual feasible region is nonempty and bounded. That is, the set $\mathcal{V} := \{(u, v) | A'u + v = c, v \geq 0\}$ is nonempty and bounded.

We note immediately that the following is a trivial consequence of the above assumption:

$$(11) \quad \mathcal{S} := \{(u, v) | A'u + v = 0, v \geq 0, (u, v) \neq 0\} = \emptyset.$$

By using a theorem of the alternative [17, Thm. 1], we have that the following is implied by (11) and hence is a consequence of Assumption 1.

LEMMA 2. Suppose that Assumption 1 holds. Then (i) the matrix A has full row rank and (ii) the set $\mathcal{X} := \{x | Ax = b, x > 0\} \neq \emptyset$.

The primal and dual least 2-norm formulations for the linear program (9) are

$$(12) \quad \min_x cx + \frac{\varepsilon}{2} xx \quad \text{s.t.} \quad Ax = b, x \geq 0$$

and

$$(13) \quad \min_{u,v} \frac{1}{2} \|A'u + v - c\|^2 - \varepsilon bu \quad \text{s.t.} \quad v \geq 0,$$

respectively, for some $\varepsilon > 0$.

If $x(\varepsilon)$ solves the primal problem (12) and $(u(\varepsilon), v(\varepsilon))$ solves the dual problem (13), then the following relation holds:

$$(14) \quad x(\varepsilon) = \frac{1}{\varepsilon} (A'u(\varepsilon) + v(\varepsilon) - c).$$

To solve problem (13) by the interior-penalty method, we solve a sequence of unconstrained subproblems

$$(15) \quad \min_{u,v} \frac{1}{2} \|A'u + v - c\|^2 - \varepsilon bu - \gamma^i \sum_{j=1}^n \log v_j,$$

where $\{\gamma^i\}$ is a sequence of decreasing positive parameters. However, for the algorithm that we are proposing here, subproblem (15) is not solved exactly. For each penalty parameter γ^i , one Newton step is taken.

Define the function $F(u, v)$ as follows:

$$F(u, v) := \frac{1}{2} \|A'u + v - c\|^2 - \varepsilon bu - \gamma^i \sum_{j=1}^n \log v_j.$$

Then its gradient and Hessian are

$$\begin{aligned} \nabla F(u, v) &= \begin{pmatrix} \nabla_u F(u, v) \\ \nabla_v F(u, v) \end{pmatrix} = \begin{pmatrix} A(A'u + v - c) - \varepsilon b \\ A'u + v - c - \gamma^i V^{-1}e \end{pmatrix}, \\ \nabla^2 F(u, v) &= \begin{pmatrix} AA' & A \\ A' & I + \gamma^i V^{-2} \end{pmatrix}, \end{aligned}$$

where $V := \text{diag}(v)$.

The Newton direction can then be obtained by solving the linear system

$$\nabla^2 F(u^i, v^i) \begin{pmatrix} u - u^i \\ v - v^i \end{pmatrix} + \nabla F(u^i, v^i) = 0$$

for u and v .

Since it is not known a priori how small ε needs be for a solution of (13) to yield the least 2-norm solution of the linear program (9), we start the algorithm with an arbitrary $\varepsilon^0 > 0$ and decrease its value as we iterate. We now state the complete algorithm.

ALGORITHM IDLN

- Initialization
 1. Choose initial points $u^0 \in \mathbb{R}^m$ and $v^0 \in \mathbb{R}_{++}^n$. Set $i = 0$.
 2. Choose initial parameters $\gamma^0 > 0$ and $\varepsilon^0 > 0$.
(A precise way to obtain these initial points and parameters is described in § 3.)
- Iteration
 1. Solve the linear system

$$(16) \quad \nabla^2 F(u^i, v^i) \begin{pmatrix} u - u^i \\ v - v^i \end{pmatrix} + \nabla F(u^i, v^i) = 0.$$

- Let (u^{i+1}, v^{i+1}) be the solution of the above linear system.
2. Set

$$(17) \quad x^{i+1} := \frac{1}{\varepsilon^i} (A'u^{i+1} + v^{i+1} - c).$$

- Termination

If the duality gap $|cx^{i+1} - bu^{i+1}|$ is sufficiently small, then stop.

Else

1. Set $i := i + 1$.

2. If $\gamma^i > \gamma_{\min}$, then $\gamma^{i+1} = \alpha\gamma^i$ for some $\alpha \in (0, 1)$.

If $\varepsilon^i > \varepsilon_{\min}$, then $\varepsilon^{i+1} = \rho\varepsilon^i$ for some $\rho \in (0, 1)$.

(The values of the attenuation factors α and ρ are given in § 3.)

3. Go to Iteration.

Remark 3. Choosing an interior point to start this algorithm is trivial, since the dual problem (13) has only nonnegativity constraints. This is the main advantage of this algorithm over the primal algorithm implemented in [10], the dual affine algorithm implemented in [23], and the primal-dual affine algorithm implemented in [13], [22], in which a Phase I is needed to start the algorithms.

Remark 4. The solution of the $m + n$ linear system (16) in the $m + n$ variables (u, v) can be achieved by first solving the m linear equations in m unknowns

$$(18) \quad \begin{aligned} & A[I - (I + \gamma(V^i)^{-2})^{-1}]A'(u - u^i) \\ & = A(I + \gamma(V^i)^{-2})^{-1}\nabla_v F(u^i, v^i) - \nabla_u F(u^i, v^i) \end{aligned}$$

for u and then computing

$$v - v^i = -(I + \gamma(V^i)^{-2})^{-1}(\nabla_v F(u^i, v^i) + A'(u - u^i)).$$

The Yale Sparse Matrix Package [6], [7] was used to solve the system of linear equations (18) for all the numerical results reported in this paper.

3. Convergence of IDLN. The logarithmic-penalty minimization problem associated with the dual problem (13) with penalty parameters $\varepsilon^i > 0$ and $\gamma^i > 0$ that we are considering is

$$(19) \quad \min_{u, v} F(u, v) := \frac{1}{2} \|A'u + v - c\|^2 - \varepsilon^i bu - \gamma^i \sum_{j=1}^n \log v_j.$$

The optimality condition for the above unconstrained problem is

$$(20) \quad A(A'u + v - c) - \varepsilon^i b = 0,$$

$$(21) \quad \gamma^i e - V(A'u + v - c) = 0,$$

where $V := \text{diag}(v)$. The Newton direction can then be obtained by solving the linear system

$$(22) \quad \begin{aligned} & \begin{pmatrix} AA' & A \\ A' & I \end{pmatrix} \begin{pmatrix} u^i \\ v^i \end{pmatrix} + \begin{pmatrix} -\varepsilon^i b - Ac \\ -c \end{pmatrix} - \begin{pmatrix} 0 \\ \gamma^i(V^i)^{-1}e \end{pmatrix} \\ & + \begin{pmatrix} AA' & A \\ A' & I + \gamma^i(V^i)^{-2} \end{pmatrix} \begin{pmatrix} u - u^i \\ v - v^i \end{pmatrix} = 0 \end{aligned}$$

or, equivalently,

$$(23) \quad A(A'u + v - c) - \varepsilon^i b = 0,$$

$$(24) \quad A'u + v - c - \gamma^i(V^i)^{-1}e + \gamma^i(V^i)^{-2}(v - v^i) = 0,$$

where $u^i \in \mathbb{R}^m$ and $v^i \in \mathbb{R}_{++}^n$. We denote the solution of the above system of linear equations by (u^{i+1}, v^{i+1}) .

Define the descent directions

$$y^i = u^{i+1} - u^i, \quad z^i = v^{i+1} - v^i,$$

and let

$$d^i = (V^i)^{-1}(v^{i+1} - v^i).$$

Premultiplying equation (24) by V^i gives

$$(25) \quad \begin{aligned} V^i(A'u^{i+1} + v^{i+1} - c) &= \gamma^i e - \gamma^i (V^i)^{-1}(v^{i+1} - v^i) \\ &= \gamma^i (e - d^i). \end{aligned}$$

Premultiplying the Newton equation (22) by the diagonal matrix

$$\begin{pmatrix} I & 0 \\ 0 & V^i \end{pmatrix}$$

gives the equation

$$\begin{aligned} &\begin{pmatrix} I & 0 \\ 0 & V^i \end{pmatrix} \begin{pmatrix} A(A'u^i + v^i - c) - \varepsilon^i b \\ A'u^i + v^i - c - \gamma^i (V^i)^{-1}e \end{pmatrix} \\ &+ \begin{pmatrix} I & 0 \\ 0 & V^i \end{pmatrix} \begin{pmatrix} AA^t & A \\ A^t & I + \gamma^i (V^i)^{-2} \end{pmatrix} \begin{pmatrix} u^{i+1} - u^i \\ v^{i+1} - v^i \end{pmatrix} = 0, \end{aligned}$$

which is equivalent to

$$(26) \quad \begin{pmatrix} A(A'u^i + v^i - c) - \varepsilon^i b \\ V^i(A'u^i + v^i - c) - \gamma^i e \end{pmatrix} + \begin{pmatrix} AA^t & AV^i \\ V^i A^t & \gamma^i I + (V^i)^2 \end{pmatrix} \begin{pmatrix} y^i \\ d^i \end{pmatrix} = 0.$$

Define the matrix M^i

$$(27) \quad M^i := \begin{pmatrix} AA^t & AV^i \\ V^i A^t & \gamma^i I + (V^i)^2 \end{pmatrix}$$

and the residual vector (p^i, r^i)

$$(28) \quad \begin{pmatrix} p^i \\ r^i \end{pmatrix} := \begin{pmatrix} \varepsilon^i b - A(A'u^i + v^i - c) \\ \gamma^i e - V^i(A'u^i + v^i - c) \end{pmatrix}.$$

Premultiplying equation (26) by (y^i, d^i) gives

$$(29) \quad \left\langle \begin{pmatrix} y^i \\ d^i \end{pmatrix}, M^i \begin{pmatrix} y^i \\ d^i \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} p^i \\ r^i \end{pmatrix}, (M^i)^{-1} \begin{pmatrix} p^i \\ r^i \end{pmatrix} \right\rangle.$$

The basic idea for the proof is as follows. Suppose that the residual vectors p^i and r^i are bounded at iteration i ; then the Newton solution (u^{i+1}, v^{i+1}) and the vector $x^{i+1} := (1/\varepsilon^i)(A'u^{i+1} + v^{i+1} - c)$ are shown to be primal-dual feasible. Moreover, by careful updating of the parameters ε and γ , the boundedness of the residual vectors p^{i+1} and r^{i+1} is guaranteed. This proof is based on the convergence proof given in [28] for the solution of a convex quadratic problem using the logarithmic-penalty method. The linear convergence of the algorithm is also established by using results given in [21]. Tseng [29] has also established the linear convergence for this algorithm.

We begin by stating the following lemmas regarding matrix M^i .

LEMMA 5. *Let M be a symmetric real $n \times n$ matrix such that, for all $x \in \mathbb{R}^n$, $\langle x, Mx \rangle \geq \gamma \|x\|^2$ for some $\gamma > 0$; then $\langle x, M^{-1}x \rangle \leq (1/\gamma) \|x\|^2$ for all $x \in \mathbb{R}^n$.*

LEMMA 6. *Let*

$$N = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix}$$

be a symmetric invertible matrix. If A^{-1} and $(C - B'A^{-1}B)^{-1}$ exist, then

$$N^{-1} := \begin{pmatrix} A^{-1} + A^{-1}B(C - B'A^{-1}B)^{-1}B'A^{-1} & -A^{-1}B(C - B'A^{-1}B)^{-1} \\ -(C - B'A^{-1}B)^{-1}B'A^{-1} & (C - B'A^{-1}B)^{-1} \end{pmatrix}.$$

LEMMA 7. Let

$$M := \begin{pmatrix} AA' & AV \\ VA' & \gamma I + V^2 \end{pmatrix},$$

where A is an $m \times n$ real matrix with independent rows, V is an $n \times n$ positive diagonal matrix, and $\gamma > 0$; then, for all $(u, v) \in \mathbb{R}^{m+n}$,

(i)

$$(30) \quad \left\langle \begin{pmatrix} u \\ v \end{pmatrix}, M \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle \geq \gamma \|v\|^2;$$

(ii)

$$(31) \quad \left\langle \begin{pmatrix} 0 \\ v \end{pmatrix}, M^{-1} \begin{pmatrix} 0 \\ v \end{pmatrix} \right\rangle \leq \frac{1}{\gamma} \|v\|^2.$$

Proof. (i) We have that

$$\left\langle \begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} AA' & AV \\ VA' & \gamma I + V^2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle = \|A'u + Vv\|^2 + \gamma \|v\|^2 \geq \gamma \|v\|^2.$$

(ii) By Lemma 6, we have that

$$\left\langle \begin{pmatrix} 0 \\ v \end{pmatrix}, M^{-1} \begin{pmatrix} 0 \\ v \end{pmatrix} \right\rangle = \langle v, N^{-1}v \rangle,$$

where

$$N^{-1} = (\gamma I + V^2 - VA'(AA')^{-1}AV)^{-1}$$

or

$$N = \gamma I + V(I - A'(AA')^{-1}A)V.$$

Define $P := (I - A'(AA')^{-1}A)$; then $P = P^2$, and we have the following:

$$\langle v, Nv \rangle = \gamma \|v\|^2 + \|PVv\|^2 \geq \gamma \|v\|^2.$$

Hence from Lemma 5 it follows that $\langle v, N^{-1}v \rangle \leq (1/\gamma)\|v\|^2$. \square

In a fashion similar to that of [12], we define the error function $E_\gamma: \mathbb{R}^m \times \mathbb{R}_+^n \rightarrow \mathbb{R}$ as

$$(32) \quad E_\gamma(u, v) := \|\gamma e - V(A'u + v - c)\|$$

to measure the error in satisfying the optimality condition (21) by the solution of the Newton equation (22). It is clear that $E_{\gamma^i}(u^{i+1}, v^{i+1}) = 0$ and $\varepsilon^i b - A(A'u^{i+1} + v^{i+1} - c) = 0$ if and only if (u^{i+1}, v^{i+1}) solves problem (19).

The next lemma gives bound to the error function E_{γ^i} at (u^{i+1}, v^{i+1}) .

LEMMA 8. Define the residual vectors p^i and r^i

$$p^i := \varepsilon^i b - A(A'u^i + v^i - c),$$

$$r^i := \gamma^i e - V^i(A'u^i + v^i - c);$$

the matrices

$$\begin{aligned} P &:= I - A'(AA')^{-1}A, \\ E^i &:= (I + (AA')^{-1}AV^i(\gamma^i I + V^i P V^i)^{-1}V^i A')(AA')^{-1}, \\ F^i &:= -(AA')^{-1}AV^i(\gamma^i I + V^i P V^i)^{-1}; \end{aligned}$$

and the scalar η^i

$$(33) \quad \eta^i := \|E^i\| \|p^i\|^2 + 2\|F^i\| \|p^i\| \|r^i\|.$$

Let (u^{i+1}, v^{i+1}) be the solution of the Newton equation (22). Then

$$E_{\gamma^i}(u^{i+1}, v^{i+1}) \leq \eta^i + \langle r^i, (\gamma^i I + V^i P V^i)^{-1} r^i \rangle.$$

Proof. Recall that

$$v^{i+1} = v^i + z^i = v^i + V^i d^i.$$

Since $V^{i+1} = \text{diag}(v^{i+1})$, then $V^{i+1} = V^i + V^i D^i$. We have the following:

$$\begin{aligned} E_{\gamma^i}(u^{i+1}, v^{i+1}) &= \|\gamma^i e - V^{i+1}(A'u^{i+1} + v^{i+1} - c)\|_2 \\ &= \|\gamma^i e - (V^i + V^i D^i)(A'u^{i+1} + v^{i+1} - c)\|_2 \\ &= \|\gamma^i e - V^i(A'u^{i+1} + v^{i+1} - c) - D^i V^i(A'u^{i+1} + v^{i+1} - c)\|_2 \\ &= \|\gamma^i e - \gamma^i e + \gamma^i d^i - D^i(\gamma^i e - \gamma^i d^i)\|_2 \quad (\text{by (25)}) \\ &= \gamma^i \|D^i d^i\|_2 \\ &\leq \gamma^i \|D^i d^i\|_1 \\ (34) \quad &= \gamma^i \|d^i\|_2^2 \\ &\leq \left\langle \begin{pmatrix} y^i \\ d^i \end{pmatrix}, M^i \begin{pmatrix} y^i \\ d^i \end{pmatrix} \right\rangle \quad (\text{by Lemma 7}) \\ &= \left\langle \begin{pmatrix} p^i \\ r^i \end{pmatrix}, (M^i)^{-1} \begin{pmatrix} p^i \\ r^i \end{pmatrix} \right\rangle \quad (\text{by (29)}) \\ &= \langle p^i, E^i p^i \rangle + 2\langle p^i, F^i r^i \rangle + \langle r^i, (\gamma^i I + V^i P V^i)^{-1} r^i \rangle \quad (\text{by Lemma 6}) \\ (35) \quad &\leq \eta^i + \langle r^i, (\gamma^i I + V^i P V^i)^{-1} r^i \rangle. \end{aligned}$$

This completes the proof. \square

For the next iteration, we update the penalty parameter

$$(36) \quad \gamma^{i+1} = \alpha \gamma^i,$$

where

$$\alpha := \frac{0.375 + \sqrt{n}}{0.5 + \sqrt{n}}.$$

We are now ready to state the following important lemma.

LEMMA 9. Let γ^{i+1} be defined as in (36) and $V^i := \text{diag}(v^i)$, where $v^i \in \mathbb{R}_{++}^n$. Define the matrix M^{i+1} and the vector r^{i+1} as

$$(37) \quad \begin{aligned} M^{i+1} &:= \begin{pmatrix} AA^t & AV^{i+1} \\ V^{i+1}A^t & \gamma^{i+1}I + (V^{i+1})^2 \end{pmatrix}, \\ r^{i+1} &:= \gamma^{i+1}e - V^{i+1}(A^t u^{i+1} + v^{i+1} - c), \end{aligned}$$

where (u^{i+1}, v^{i+1}) is the solution of the Newton equation (22). Suppose that

$$(38) \quad \langle r^i, (\gamma^i I + V^i P V^i)^{-1} r^i \rangle \leq 0.25 \gamma^i$$

and that η^i as defined in (33) satisfies

$$(39) \quad \eta^i \leq 0.125 \gamma^i.$$

Then

(i) The point (u^{i+1}, v^{i+1}) is feasible for the dual problem (13), $v^{i+1} > 0$, $x^{i+1} := (1/\varepsilon^i)(A^t u^{i+1} + v^{i+1} - c)$ is feasible for the primal problem (12) with $\varepsilon = \varepsilon^i$, and the following holds:

$$(40) \quad \varepsilon^i x_j^{i+1} v_j^{i+1} \leq \gamma^i \quad \forall j = 1, 2, \dots, n.$$

(ii) The vector r^{i+1} is bounded as follows:

$$\langle r^{i+1}, (\gamma^{i+1} I + V^{i+1} P V^{i+1})^{-1} r^{i+1} \rangle \leq 0.25 \gamma^{i+1}.$$

Proof. (i) We will first show that under the above conditions $\|d^i\| < 1$, where $d^i = (V^i)^{-1}(v^{i+1} - v^i)$. By (34) and (35) of the proof of Lemma 8, we have

$$\begin{aligned} \|d^i\|^2 &\leq \frac{1}{\gamma^i} (\eta^i + \langle r^i, (\gamma^i I + V^i P V^i)^{-1} r^i \rangle) \\ &< 1. \end{aligned}$$

The fact that $\|d^i\| < 1$ and $v^i > 0$ implies that $v^{i+1} > 0$ and hence the dual feasibility of (u^{i+1}, v^{i+1}) . From (25) and the definition of x^{i+1} , we have

$$\begin{aligned} \varepsilon^i x^{i+1} &= A^t u^{i+1} + v^{i+1} - c \\ &= \gamma^i (V^i)^{-1} (e - d^i) > 0. \end{aligned}$$

The equality constraint $Ax^{i+1} = b$ follows from the definition of x^{i+1} and the Newton equation (23). To establish relation (40), note that from the definition $d^i = (V^i)^{-1}(v^{i+1} - v^i)$ we have

$$\begin{aligned} \varepsilon^i x^{i+1} &= \gamma^i (V^i)^{-1} (e - d^i) \\ &= \gamma^i (V^{i+1})^{-1} (D^i + I) (e - d^i) \\ &= \gamma^i (V^{i+1})^{-1} (e - D^i d^i) \\ &\leq \gamma^i (V^{i+1})^{-1} e. \end{aligned}$$

Upon premultiplying the last relation by V^{i+1} , we get $\varepsilon^i x_j^{i+1} v_j^{i+1} \leq \gamma^i$ for all $j = 1, 2, \dots, n$.

(ii) The proof of the second part of Lemma 9 is as follows:

$$\begin{aligned}
 & \left[\frac{1}{\gamma^{i+1}} \langle r^{i+1}, (\gamma^{i+1}I + V^{i+1}PV^{i+1})^{-1}r^{i+1} \rangle \right]^{1/2} \\
 &= \left[\frac{1}{\gamma^{i+1}} \left\langle \begin{pmatrix} 0 \\ r^{i+1} \end{pmatrix}, (M^{i+1})^{-1} \begin{pmatrix} 0 \\ r^{i+1} \end{pmatrix} \right\rangle \right]^{1/2} \\
 &\leq \frac{1}{\gamma^{i+1}} \|r^{i+1}\| \quad (\text{by Lemma 7}) \\
 &= \frac{1}{\gamma^{i+1}} \|\gamma^{i+1}e - V^{i+1}(A^t u^{i+1} + v^{i+1} - c)\| \\
 &= \frac{1}{\alpha \gamma^i} \|\alpha \gamma^i e - V^{i+1}(A^t u^{i+1} + v^{i+1} - c)\| \quad (\text{definition of } \gamma^{i+1}) \\
 &\leq \frac{1}{\alpha \gamma^i} (E_{\gamma^i}(u^{i+1}, v^{i+1}) + (1 - \alpha)\gamma^i \|e\|) \\
 &\leq \frac{1}{\alpha \gamma^i} (\eta^i + \langle r^i, (\gamma^i I + V^i P V^i)^{-1} r^i \rangle) + \frac{1 - \alpha}{\alpha} \|e\| \quad (\text{by Lemma 8}) \\
 &\leq \frac{1}{\alpha} (0.125 + 0.25 + \sqrt{n}) - \sqrt{n} \\
 &\leq 0.5. \quad \square
 \end{aligned}$$

The next two lemmas establish the boundedness of u^{i+1} , v^{i+1} , and x^{i+1} under Assumption 1. We will show that, if the conditions (38) and (39) of Lemma 9 are satisfied, then x^{i+1} is bounded. The proof is similar to that of Polyak in [26] for the gradient-projection algorithm. The boundedness of x^{i+1} and the assumption that the dual feasible set is bounded establish the boundedness of (u^{i+1}, v^{i+1}) .

LEMMA 10. *Suppose that the conditions (38) and (39) of Lemma 9 are satisfied by $(u^i, v^i) \in \mathbb{R}^m \times \mathbb{R}_{++}^n$. Let (u^{i+1}, v^{i+1}) be the solution of the Newton equation (22), and let x^* be a solution of the linear program (9). If the parameters γ^i and ε^i are such that $\gamma^i \leq (\varepsilon^i)^2$, then*

$$(41) \quad \|x^{i+1} - x^*\|^2 \leq 2n + \|x^*\|^2,$$

where $x^{i+1} = (1/\varepsilon^i)(A^t u^{i+1} + v^{i+1} - c)$.

Proof. From the Newton equation (23), we have

$$A(A^t u^{i+1} + v^{i+1} - c) = \varepsilon^i b,$$

which gives

$$u^{i+1} = (AA^t)^{-1}(\varepsilon^i b - A(v^{i+1} - c)).$$

Hence

$$\begin{aligned}
 (42) \quad x^{i+1} &= (I - A'(AA^t)^{-1}A) \left(\frac{1}{\varepsilon^i} (v^{i+1} - c) \right) + A'(AA^t)^{-1}b \\
 &= P_Q \left(\frac{1}{\varepsilon^i} (v^{i+1} - c) \right),
 \end{aligned}$$

where $P_Q(x)$ is the projection of x onto the set $Q := \{z \mid Az = b\}$. By the minimum principle [26, p. 121] applied to the above projection problem (42), we have

$$(43) \quad 0 \geq \left\langle \frac{1}{\varepsilon^i} (v^{i+1} - c) - x^{i+1}, x^* - x^{i+1} \right\rangle$$

or, equivalently,

$$(44) \quad \begin{aligned} 0 &\geq \langle v^{i+1} - c - \varepsilon^i x^{i+1}, x^* - x^{i+1} \rangle \\ &= -\langle c, x^* - x^{i+1} \rangle + \langle v^{i+1}, x^* \rangle - \langle v^{i+1}, x^{i+1} \rangle - \varepsilon^i \langle x^{i+1}, x^* - x^{i+1} \rangle \\ &\geq \varepsilon^i \langle -x^{i+1}, x^* - x^{i+1} \rangle - n \frac{\gamma^i}{\varepsilon^i} \\ &= \frac{1}{2} \varepsilon^i (\|x^{i+1}\|^2 + \|x^{i+1} - x^*\|^2 - \|x^*\|^2) - n \frac{\gamma^i}{\varepsilon^i}. \end{aligned}$$

The second inequality in (44) follows from the fact that $cx^* \leq cx^{i+1}$, $\langle v^{i+1}, x^* \rangle \geq 0$, and $\varepsilon^i x_j^{i+1} v_j^{i+1} \leq \gamma^i$ for all $j = 1, 2, \dots, n$. Rearranging the terms in (44) and multiplying by $2/\varepsilon^i$ gives

$$\begin{aligned} \|x^{i+1} - x^*\|^2 &\leq 2n \frac{\gamma^i}{(\varepsilon^i)^2} + \|x^*\|^2 - \|x^{i+1}\|^2 \\ &\leq 2n + \|x^*\|^2. \end{aligned}$$

Hence the proof is complete. \square

In the next lemma, we establish the boundedness of (u^{i+1}, v^{i+1}) .

LEMMA 11. *Suppose that the point $(u^i, v^i) \in \mathbb{R}^m \times \mathbb{R}_{++}^n$ satisfies the conditions (38) and (39) of Lemma 9. Let (u^{i+1}, v^{i+1}) be the solution of the Newton equation (22) and $\gamma^i \leq (\varepsilon^i)^2$. Furthermore, suppose that the set $\mathcal{V} := \{(u, v) \mid A'u + v = c, v \geq 0\}$ is bounded. Then there exists a constant $\tau < \infty$ depending only on the matrix A and vectors b and c of the linear program (9) such that*

$$(45) \quad \|u^{i+1}, v^{i+1}\| \leq \tau.$$

Proof. Define the set \mathcal{V}^i as follows:

$$(46) \quad \mathcal{V}^i := \{(u, v) \mid A'u + v = c + \varepsilon^i x^{i+1}, v \geq 0\}.$$

Note that \mathcal{V}^i is nonempty by the construction of $x^{i+1} = (1/\varepsilon^i)(A'u^{i+1} + v^{i+1} - c)$. We claim that the set \mathcal{V}^i is bounded. In Lemma 10 it was shown that x^{i+1} is bounded. Hence, if the set \mathcal{V}^i is unbounded, then there exists (\bar{u}, \bar{v}) such that

$$A'\bar{u} + \bar{v} = 0, \quad \bar{v} \geq 0, \quad (\bar{u}, \bar{v}) \neq 0.$$

Then for any point $(w, z) \in \mathcal{V}$ we have that $(w + \lambda \bar{u}, z + \lambda \bar{v}) \in \mathcal{V}$ for any $\lambda \geq 0$, which contradicts the assumption that the set \mathcal{V} is bounded. Hence \mathcal{V}^i is bounded.

Consider now the following nonconvex problem:

$$(47) \quad \max_{u, v} \|(u, v)\| \quad \text{s.t. } A'u + v = c + \varepsilon^i x^{i+1}, v \geq 0.$$

This problem has a solution, since we have just shown that its feasible set is bounded. By the generalized theorem of the existence of a basic feasible solution [20], it follows that there must exist a basic solution. Let the basis matrix B^i denote the $n \times n$

nonsingular submatrix of $[A' \ I]$ corresponding to the basic solution (\tilde{u}, \tilde{v}) of problem (47). We have

$$\begin{aligned}\|\tilde{u}, \tilde{v}\| &= \|(B^i)^{-1}(c + \varepsilon^i x^{i+1})\| \\ &\leq \|(B^i)^{-1}\| \|c + \varepsilon^i x^{i+1}\|.\end{aligned}$$

Since there are only a finite number of basis matrices in $[A' \ I]$ and since both ε^i and x^{i+1} are bounded, we conclude that there must exist $\tau < \infty$ such that

$$\|u^{i+1}, v^{i+1}\| \leq \|\tilde{u}, \tilde{v}\| \leq \tau,$$

and this completes the proof. \square

From Lemma 11 we have that both $\max_v \|AV\|$ and $\max_v \|VA'\|$ such that $v \in \mathcal{V}^i$ and $V := \text{diag}(v)$ are finite, where \mathcal{V}^i is the set defined by (46).

The next lemma shows that, if the attenuation factor $\rho \in (0, 1)$ for decreasing ε^i is chosen carefully, then the assumption (39) of Lemma 9 holds at iteration $i+1$.

LEMMA 12. Let (u^{i+1}, v^{i+1}) be the solution of the Newton equation (22) and $x^{i+1} = (1/\varepsilon^{i+1})(A'u^{i+1} + v^{i+1} - c)$. Suppose that $(u^i, v^i) \in \mathbb{R}^m \times \mathbb{R}_{++}^n$ satisfy the conditions (38) and (39) of Lemma 9 and that the sequences $\{\gamma^k\}$ and $\{\varepsilon^k\}$ are such that

$$(48) \quad 0 < \{\gamma^k\} \leq \gamma_{\max}$$

and

$$(49) \quad 0 < \{\varepsilon^k\} \leq \varepsilon_{\max}.$$

Define the constants

$$K_1 = \|(AA')^{-1}\|,$$

$$K_2^i = \max \left\{ \max_v \|AV\|, \max_v \|VA'\| \right\} \quad \text{s.t. } A'u + v = c + \varepsilon^i x^{i+1}, v \geq 0,$$

$$C_1^i = (\gamma_{\max} + K_1(K_2^i)^2)K_1\varepsilon_{\max}^2 \|b\|^2 / \gamma^{i+1},$$

$$C_2^i = (4/\alpha)\varepsilon_{\max}\sqrt{n} \|b\| K_1 K_2^i,$$

where $\alpha = (0.375 + \sqrt{n})/(0.5 + \sqrt{n})$ and $\gamma^{i+1} = \alpha\gamma^i$. If

$$\varepsilon^{i+1} = \rho^i \varepsilon^i,$$

where

$$(50) \quad 1 > \rho^i \geq 1 - \delta^i$$

and

$$(51) \quad 0 < \delta^i \leq (-C_2^i + \sqrt{(C_2^i)^2 + 0.5\gamma^{i+1}C_1^i})/2C_1^i,$$

then we have

$$\eta^{i+1} \leq 0.125\gamma^{i+1},$$

where

$$\eta^{i+1} := \|E^{i+1}\| \|p^{i+1}\|^2 + 2\|F^{i+1}\| \|p^{i+1}\| \|r^{i+1}\|,$$

$$p^{i+1} := \varepsilon^{i+1}b - A(A'u^{i+1} + v^{i+1} - c),$$

$$r^{i+1} := \gamma^{i+1}e - V^{i+1}(A'u^{i+1} + v^{i+1} - c),$$

$$E^{i+1} := (I + (AA')^{-1}AV^{i+1}(\gamma^{i+1}I + V^{i+1}PV^{i+1})^{-1}V^{i+1}A')(AA')^{-1},$$

$$F^{i+1} := -(AA')^{-1}AV^{i+1}(\gamma^{i+1}I + V^{i+1}PV^{i+1})^{-1}.$$

Proof. We will first compute the bounds on the norms of the residual vectors p^{i+1} and r^{i+1} as follows:

(i)

$$\begin{aligned}\|p^{i+1}\| &= \|\varepsilon^{i+1}b - A(A'u^{i+1} + v^{i+1} - c)\| \\ &= \|\varepsilon^i b - A(A'u^{i+1} + v^{i+1} - c) + \varepsilon^{i+1}b - \varepsilon^i b\| \\ &= (1 - \rho^i)\varepsilon^i \|b\| \quad (\text{by (23)}),\end{aligned}$$

(ii)

$$\begin{aligned}\|r^{i+1}\| &= \|\gamma^{i+1}e - V^{i+1}(A'u^{i+1} + v^{i+1} - c)\| \\ &\leq \gamma^{i+1}\sqrt{n} + \|\varepsilon^i V^{i+1}x^{i+1}\| \\ &\leq 2\gamma^i\sqrt{n} \quad (\text{by (40)}).\end{aligned}$$

Next we compute the bounds on the norm of the matrices E^{i+1} and F^{i+1} as follows:

(i)

$$\begin{aligned}\|E^{i+1}\| &= \|(I + (AA')^{-1}AV^{i+1}(\gamma^{i+1}I + V^{i+1}PV^{i+1})^{-1}V^{i+1}A')(AA')^{-1}\| \\ &\leq \left(1 + \frac{1}{\gamma^{i+1}}K_1(K_2^i)^2\right)K_1,\end{aligned}$$

(ii)

$$\begin{aligned}\|F^{i+1}\| &= \|-(AA')^{-1}AV^{i+1}(\gamma^{i+1}I + V^{i+1}PV^{i+1})^{-1}\| \\ &\leq \frac{1}{\gamma^{i+1}}K_1K_2^i.\end{aligned}$$

Hence we have

$$\begin{aligned}\eta^{i+1} &= \|E^{i+1}\| \|p^{i+1}\|^2 + 2\|F^{i+1}\| \|p^{i+1}\| \|r^{i+1}\| \\ &\leq \frac{1}{\gamma^{i+1}}(\gamma^{i+1} + K_1(K_2^i)^2)K_1(1 - \rho^i)^2(\varepsilon^i)^2\|b\|^2 + 2\frac{1}{\gamma^{i+1}}K_1K_2^i(1 - \rho^i)\varepsilon^i\|b\|2\gamma^i\sqrt{n} \\ &\leq \varepsilon_{\max}^2(\gamma_{\max} + K_1(K_2^i)^2)K_1\|b\|^2(1 - \rho^i)^2/\gamma^{i+1} + (4/\alpha)\varepsilon_{\max}K_1K_2^i\|b\|\sqrt{n}(1 - \rho^i) \\ &\leq C_1^i(1 - \rho^i)^2 + C_2^i(1 - \rho^i) \\ &\leq C_1^i(\delta^i)^2 + C_2^i(\delta^i) \quad (\text{by (50)}) \\ &\leq 0.125\gamma^{i+1} \quad (\text{by (51)}).\end{aligned}$$

This completes the proof of the lemma. \square

By using the results from the Lemmas 9–12, we can now establish the following theorem regarding Algorithm IDLN.

THEOREM 13. Let $(u^i, v^i) \in \mathbb{R}^m \times \mathbb{R}_{++}^n$ be the i th iterate of Algorithm IDLN with parameter $\varepsilon = \varepsilon^i$ and $\gamma = \gamma^i$, $\gamma^i \leq (\varepsilon^i)^2$ such that the following two conditions are satisfied:

$$(52) \quad \langle r^i, (\gamma^i I + V^i P V^i)^{-1} r^i \rangle \leq 0.25\gamma^i$$

and

$$(53) \quad \eta^i \leq 0.125\gamma^i,$$

where r^i is the residual vector defined by (28) and η^i is the real number defined by (33) in Lemma 8. Suppose that (u^{i+1}, v^{i+1}) is the solution of the Newton equation (22). If we let

$$x^{i+1} = \frac{1}{\varepsilon^i} (A^t u^{i+1} + v^{i+1} - c),$$

$$\gamma^{i+1} = \alpha \gamma^i,$$

$$\varepsilon^{i+1} = \rho^i \varepsilon^i,$$

where $\alpha = (0.375 + \sqrt{n}) / (0.5 + \sqrt{n})$ and $\rho^i \in (0, 1)$, satisfying condition (50) in Lemma 12, then

(i) The triple $(x^{i+1}, u^{i+1}, v^{i+1})$ is bounded and is feasible for the primal-dual problems (12), (13) with $\varepsilon = \varepsilon^i$, $v^{i+1} > 0$, and

$$\varepsilon^i x_j^{i+1} v_j^{i+1} \leq \gamma^i \quad \forall j = 1, 2, \dots, n$$

holds, and

(ii) The bounds

$$\eta^{i+1} \leq 0.125 \gamma^{i+1}$$

and

$$\langle r^{i+1}, (\gamma^{i+1} + V^{i+1} P V^{i+1})^{-1} r^{i+1} \rangle \leq 0.25 \gamma^{i+1}$$

are satisfied for (u^{i+1}, v^{i+1}) .

The idea for the linear convergence proof of IDLN comes from a proof given by Mangasarian and De Leone [21] for the least 2-norm solution of linear programs, in which they give error bounds for a class of more general problems. The problem they consider is

$$(54) \quad \min_x f(x) \quad \text{s.t. } x \in S := \{x \mid x \geq 0, g(x) \leq 0\}.$$

We begin by restating their main result.

THEOREM 14 (see [21, Thm. 2.2]). *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ be differentiable on \mathbb{R}^n , let f be strongly convex on \mathbb{R}^n with positive constant k , and let g be convex on \mathbb{R}^n . Let g be linear and $S \neq \emptyset$ or let g satisfy the Slater constraint qualification, that is,*

$$g(\hat{x}) < 0, \quad \hat{x} > 0,$$

for some $\hat{x} \in \mathbb{R}^n$. Then for any $(x, u) \in \mathbb{R}^n \times \mathbb{R}_+^m$ the distance $\|x - \bar{x}\|$ to the unique solution \bar{x} of (54) is bounded by

$$\begin{aligned} k^{1/2} \|x - \bar{x}\| &\leq [x \nabla_x L(x, u) - u g(x) + \alpha \|(-\nabla_x L(x, u))_+\|_1 \\ &\quad + \beta \|g(x)_+\|_\infty + \gamma \|(-x)_+\|_\infty]^{1/2}, \end{aligned}$$

where

$$L(x, u) := f(x) + u g(x),$$

$$\alpha := \min_{x \in S} (\|x\|_\infty + \|\nabla f(x)\|_1 / k),$$

$$\beta := \min_{(u, v) \in W} \|u\|_1,$$

$$\gamma := \min_{(u, v) \in W} \|v\|_1,$$

where $W \subset \mathbb{R}_+^{m+n}$ is the nonempty closed convex polyhedral set of optimal multipliers (u, v) of the convex program (54) associated with the constraints $g(x) \leq 0$, $x \geq 0$.

By using Theorem 14 we will show that if we impose a stronger condition on the parameter ε^{i+1} , then IDLN Algorithm is linearly convergent.

THEOREM 15. *Let $(u^{i+1}, v^{i+1}, x^{i+1})$, α , and ρ^i be as in Theorem 13. Suppose that all the conditions in Theorem 13 are satisfied, and suppose that the parameter ε is decreased as*

$$(55) \quad \varepsilon^{i+1} = \bar{\rho}^i \varepsilon^i,$$

where $1 > \bar{\rho}^i := \max\{\alpha^{1/4}, \rho^i\}$, ρ^i as defined by (50), and that γ is decreased as

$$(56) \quad \gamma^{i+1} = \alpha \gamma^i,$$

where $\alpha = (0.375 + \sqrt{n}) / (0.5 + \sqrt{n})$. Then the sequence $\{x^i\}$ converges to \bar{x} , the unique least 2-norm solution of (9) with the linear root rate [25]

$$(57) \quad \|x^{i+1} - \bar{x}\| \leq \delta(\alpha^{1/4})^{i+1} \quad \text{for } i \geq \bar{i}$$

for some constant δ and some integer \bar{i} .

Proof. Let $L(x, u) = cx + (\varepsilon^i/2)xx - u'(Ax - b)$. Then

$$\nabla_x L(x^{i+1}, u^{i+1}) = c + \varepsilon^i x^{i+1} - A^t u^{i+1} = v^{i+1} > 0.$$

By Theorem 13 we have that

$$\begin{aligned} v^{i+1} &\geq 0, \\ x^{i+1} &\geq 0, \\ \varepsilon^i x_j^{i+1} v_j^{i+1} &\leq \gamma^i \quad \forall j = 1, 2, \dots, n, \\ Ax^{i+1} &= b. \end{aligned}$$

Let $\bar{x}(\varepsilon^i)$ be the solution of the quadratic problem (12) with $\varepsilon = \varepsilon^i$. It follows from Theorem 14 that

$$\begin{aligned} \|x^{i+1} - \bar{x}(\varepsilon^i)\| &= \frac{1}{\sqrt{\varepsilon^i}} \langle x^{i+1}, v^{i+1} \rangle^{1/2} \\ &\leq \frac{1}{\sqrt{\varepsilon^i}} (n\gamma^i / \varepsilon^i)^{1/2} \\ &= (n\gamma^i)^{1/2} / \varepsilon^i \\ &\leq (n(\alpha)^i \gamma^0)^{1/2} / (\alpha^{1/4})^i \varepsilon^0 \\ &= \delta(\alpha^{1/4})^{i+1}, \end{aligned}$$

where $\delta = \sqrt{n\gamma^0} / (\varepsilon^0 \alpha^{1/4})$. Now let \bar{i} be the smallest integer such that $\varepsilon^{\bar{i}} \leq \bar{\varepsilon}$, where $\bar{\varepsilon}$ is that defined below (2). Combining the last result and the fact that $\bar{x} = \bar{x}(\varepsilon^i)$ for $i \geq \bar{i}$, we have

$$\begin{aligned} \|x^{i+1} - \bar{x}\| &\leq \|x^{i+1} - \bar{x}(\varepsilon^i)\| + \|\bar{x}(\varepsilon^i) - \bar{x}\| \\ &= \|x^{i+1} - \bar{x}(\varepsilon^i)\| \\ &\leq \delta(\alpha^{1/4})^{i+1}. \end{aligned}$$

This establishes the linear convergence of the iterates. \square

Remark 16. The condition $\gamma^i \leq (\varepsilon^i)^2$ required in Theorem 13 will be satisfied for all i if we let $\gamma^0 = (\varepsilon^0)^2$ and if $\{\varepsilon^i\}$ and $\{\gamma^i\}$ are decreased according to (55) and (56).

Remark 17. The parameter ε^k in (49) need not go to zero. Let \bar{i} be the smallest integer such that $\varepsilon^{\bar{i}} \leq \bar{\varepsilon}$, where $\bar{\varepsilon}$ is defined below expression (2). If for all $k > \bar{i}$ we fix $\varepsilon^k = \bar{\varepsilon}$, then the linear convergence of the algorithm still holds.

Remark 18. Suppose that constraint matrix A and the right-hand-side vector b of the linear program (9) have the form

$$A = \begin{pmatrix} \bar{A} \\ e' \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ n \end{pmatrix},$$

where \bar{A} is some $(m-1) \times n$ matrix (see [4] for details on how to transform the constraints of a general linear program into this form). It is also assumed that the point e satisfies $Ae = b$. We claim that an initial feasible point satisfying the conditions (52) and (53) of Theorem 13 can be found immediately. To see this, define $\gamma^0 = (\varepsilon^0)^2$ for some positive ε^0 such that $\|(I - A'(AA')^{-1}A)c\| \leq 0.5\varepsilon^0$. Let $u^0 = (AA')^{-1}Ac$ and let $v^0 = \varepsilon^0 e$. We have

TABLE 1
Dimensions of linear problems.

Pr. No.	Problem Name	Original			Adjusted		
		rows	columns	nonzeros	rows	columns	nonzeros
1	25fv47	822	1571	11127	820	1876	10705
2	Adlittle	57	97	465	56	138	424
3	Afiro	28	32	88	27	51	102
4	Agg	489	163	2541	488	615	2862
5	Agg2	517	302	4515	516	758	4750
6	Agg3	517	302	4531	516	758	4756
7	Bandm	306	472	2659	305	472	2494
8	Beaconfd	174	262	3476	173	295	3408
9	Blend	75	83	521	74	114	522
10	Bnl1	644	1175	6129	642	1586	5532
11	Bnl2	2325	3489	16124	2324	4486	14996
12	Bore3d	234	315	1525	246	346	1473
13	Brandy	221	249	2150	193	303	2202
14	Capri	272	353	1786	446	641	2230
15	Cre-a	3517	4067	19054	3428	7248	18168
16	Cre-c	3069	3678	16922	2986	6411	15977
17	Czprob	930	3523	14173	1158	3562	10937
18	D2q06c	2172	5167	35674	2171	5831	33081
19	Degen2	445	534	4449	444	757	4201
20	Degen3	1504	1818	26230	1503	2604	25432
21	E226	224	282	2767	223	472	2768
22	Ffff800	525	854	6235	524	1028	6401
23	Finnis	498	614	2714	619	1141	2959
24	Gfrd-pnc	617	1092	3467	876	1420	2965
25	Grow15	301	645	5665	900	1245	6820
26	Grow22	441	946	8318	1320	1826	10012
27	Grow7	141	301	2633	420	581	3172
28	Israel	175	142	2358	174	316	2443
29	Kb2	44	41	291	52	77	331
30	Lotfi	154	308	1086	153	366	1136
31	Pilot.we	723	2789	9218	1256	3384	10255
32	Rabo	391	576	5510	317	560	5201
33	Recipe	92	180	752	211	300	903

TABLE 2
Dimensions of linear problems (continued).

Pr. No.	Problem Name	Original			Adjusted		
		rows	columns	nonzeros	rows	columns	nonzeros
34	Sc105	106	103	281	105	163	340
35	Sc205	206	203	552	205	317	665
36	Sc50a	51	48	131	50	78	160
37	Sc50b	51	48	119	50	78	148
38	Scagr25	472	500	2029	471	671	1725
39	Scagr7	130	140	553	129	185	465
40	Scfxm1	331	457	2612	330	600	2732
41	Scfxm2	661	914	5229	660	1200	5469
42	Scfxm3	991	1371	7846	990	1800	8206
43	Scorpion	389	358	1708	388	466	1534
44	Scrs8	491	1169	4029	490	1275	3288
45	Scsd1	78	760	3148	77	760	2388
46	Scsd6	148	1350	5666	147	1350	4316
47	Scsd8	398	2750	11334	397	2750	8584
48	Sctap1	301	480	2052	300	660	1872
49	Sctap2	1091	1880	8124	1090	2500	7334
50	Sctap3	1481	2480	10734	1480	3340	9734
51	Share1b	118	225	1182	117	253	1179
52	Share2b	97	79	730	96	162	777
53	Ship041	403	2118	8450	360	2166	6380
54	Ship04s	403	1458	5810	360	1506	4400
55	Ship081	779	4283	17085	712	4363	12882
56	Ship08s	779	2387	9501	712	2467	7194
57	Ship121	1152	5427	21597	1042	5533	16276
58	Ship12s	1152	2763	10941	1042	2869	8284
59	Stocfor1	118	111	474	117	165	501
60	Stocfor2	2158	2031	9492	2157	3045	9357
61	Truss1	201	1602	6586	200	1602	4984
62	Truss2	501	4312	17896	500	4312	13584
63	Truss3	1001	8806	36642	1000	8806	27836
64	Vtp.base	199	203	914	347	477	1331
65	Wood1p	245	2594	70216	244	2595	70216
66	Woodw	1099	8405	37478	1098	8418	37487

(i)

$$\begin{aligned}
 p^0 &= \varepsilon^0 b - A(A'^u + v^0 - c) \\
 &= \varepsilon^0 b - A(A'(AA')^{-1}Ac + v^0 - c) \\
 &= \varepsilon^0 b - Av^0 \\
 &= \varepsilon^0 b - \varepsilon^0 Ae \\
 &= 0,
 \end{aligned}$$

(ii)

$$\begin{aligned}
 r^0 &= \gamma^0 e - V^0(A'^u + v^0 - c) \\
 &= \gamma^0 e - V^0(A'(AA')^{-1}Ac + v^0 - c) \\
 &= \gamma^0 e + V^0(I - A'(AA')^{-1}A)c - V^0 v^0 \\
 &= \gamma^0 e + V^0(I - A'(AA')^{-1}A)c - (\varepsilon^0)^2 e \\
 &= V^0(I - A'(AA')^{-1}A)c.
 \end{aligned}$$

TABLE 3
 IDLN results.

Pr. No.	Problem Name	Primal Infeasibility	Dual Infeasibility	Duality Gap	Comple-mentarity
1	25fv47	5.38E-11	1.38E-14	1.76E-13	2.21E-17
2	Adlittle	4.12E-14	2.13E-12	3.23E-16	5.95E-15
3	Afiro	2.86E-11	1.40E-09	7.95E-16	1.48E-11
4	Agg	4.17E-17	1.80E-12	1.43E-13	6.13E-18
5	Agg2	1.55E-16	8.15E-16	3.52E-14	2.07E-16
6	Agg3	1.27E-14	2.15E-11	2.00E-10	4.34E-13
7	BandM	2.74E-15	5.87E-16	1.21E-12	2.45E-18
8	Beaconfd	8.33E-15	1.41E-15	3.25E-15	3.78E-20
9	Blend	7.23E-13	3.81E-11	1.54E-14	8.19E-14
10	Bnl1	1.14E-11	0.00E+00	1.16E-06	2.00E-12
11	Bnl2	7.32E-12	1.80E-05	1.89E-13	2.82E-18
12	Bore3d	6.42E-14	4.45E-12	7.58E-14	2.58E-20
13	BrandY	6.64E-14	5.48E-15	5.76E-15	1.00E-20
14	Capri	1.74E-16	1.87E-10	9.83E-13	2.51E-21
15	Cre-a	6.11E-14	2.81E-05	4.90E-12	3.44E-16
16	Cre-c	3.39E-12	9.31E-07	1.54E-07	7.76E-10
17	CzProb	1.33E-14	1.27E-10	1.12E-13	6.95E-18
18	D2q06c	5.85E-13	1.48E-08	1.13E-09	1.68E-11
19	Degen2	2.38E-11	1.81E-11	1.74E-11	2.04E-17
20	Degen3	3.65E-11	9.03E-12	1.81E-12	6.78E-18
21	E226	7.62E-11	2.80E-11	1.37E-10	1.16E-17
22	Ffff800	5.03E-16	2.77E-06	3.14E-07	9.14E-16
23	Finnis	7.78E-13	1.93E-05	6.42E-08	2.45E-09
24	Gfrd-Pnc	2.51E-14	7.59E-11	7.42E-15	2.26E-20
25	Grow15	1.35E-16	1.72E-15	0.00E+00	1.18E-18
26	Grow22	1.37E-16	2.50E-15	1.43E-13	3.25E-15
27	Grow7	1.40E-16	2.70E-16	1.56E-16	2.47E-18
28	Israel	2.28E-16	1.33E-11	1.83E-06	4.75E-10
29	Kb2	1.09E-14	1.53E-16	1.65E-13	7.51E-16
30	Lotfi	2.79E-14	9.28E-17	5.51E-13	8.10E-17
31	Pilot.we	4.61E-16	1.32E-11	6.72E-06	1.32E-13
32	Rabo	1.61E-16	6.77E-15	6.56E-16	1.01E-18
33	Recipe	1.13E-14	2.44E-14	0.00E+00	1.64E-23

Hence

$$\begin{aligned}
 \|r^0\| &= \|V^0(I - A'(AA')^{-1}A)c\| \\
 &= \varepsilon^0\|(I - A'(AA')^{-1}A)c\| \\
 &\leq 0.5(\varepsilon^0)^2 \\
 &= 0.5\gamma^0.
 \end{aligned}$$

Since $p^0 = 0$, we have that

$$\eta^0 = 0$$

and

$$\begin{aligned}
 \langle r^0, (\gamma^0 I + V^0 P V^0)^{-1} r^0 \rangle &\leq \frac{1}{\gamma^0} \|r^0\|^2 \\
 &\leq \frac{1}{\gamma^0} \frac{(\gamma^0)^2}{4} \\
 &= 0.25\gamma^0.
 \end{aligned}$$

TABLE 4
IDLN results (continued).

Pr. No.	Problem Name	Primal Infeasibility	Dual Infeasibility	Duality Gap	Complementarity
34	Sc105	3.85E-13	1.14E-16	7.75E-12	5.50E-19
35	Sc205	1.32E-12	8.48E-17	8.06E-14	6.56E-21
36	Sc50a	4.23E-11	3.28E-17	4.76E-11	8.62E-19
37	Sc50b	7.66E-12	5.65E-10	6.09E-16	1.01E-17
38	Scagr25	2.47E-13	8.18E-14	6.95E-13	1.39E-14
39	Scagr7	2.43E-14	8.17E-10	4.45E-10	1.24E-11
40	Scfxm1	2.10E-12	6.81E-08	1.38E-14	5.21E-17
41	Scfxm2	4.72E-14	2.88E-08	5.61E-11	2.14E-15
42	Scfxm3	3.09E-14	1.78E-09	2.42E-13	6.14E-18
43	Scorpion	2.55E-11	3.68E-10	7.77E-12	3.43E-20
44	Scrs8	2.24E-14	8.26E-15	1.35E-14	9.01E-21
45	ScSd1	5.95E-10	1.85E-07	2.73E-10	8.52E-11
46	ScSd6	1.07E-11	5.81E-08	2.24E-12	1.50E-13
47	ScSd8	4.86E-12	1.05E-07	4.42E-13	1.35E-13
48	ScTap1	1.44E-13	1.10E-12	2.29E-13	1.09E-17
49	ScTap2	2.95E-14	1.82E-12	3.16E-15	3.48E-17
50	ScTap3	1.10E-12	3.71E-12	2.27E-14	3.25E-16
51	Share1b	2.83E-14	1.70E-15	2.78E-14	2.56E-17
52	Share2b	1.67E-11	5.74E-15	2.50E-13	6.66E-18
53	Ship041	1.30E-14	7.95E-12	1.72E-14	3.18E-16
54	Ship04s	1.05E-11	1.22E-11	1.41E-12	7.03E-19
55	Ship081	3.16E-13	4.03E-10	3.21E-13	5.56E-15
56	Ship08s	8.48E-14	8.06E-12	1.57E-13	3.07E-15
57	Ship121	1.24E-10	2.24E-09	1.64E-13	4.51E-18
58	Ship12s	8.84E-13	6.15E-12	1.59E-13	1.03E-15
59	Stocfor1	8.02E-12	7.50E-14	1.34E-10	1.24E-13
60	Stocfor2	4.95E-10	1.87E-13	1.99E-14	3.23E-17
61	Truss1	1.61E-14	1.26E-10	2.07E-15	8.07E-17
62	Truss2	4.34E-12	2.74E-13	1.32E-13	1.11E-18
63	Truss3	2.75E-09	2.30E-12	1.27E-10	8.81E-16
64	Vtp.base	8.97E-16	2.04E-10	6.72E-16	1.21E-22
65	Wood1p	6.92E-07	2.64E-08	4.25E-09	2.25E-15
66	Woodw	3.27E-05	0.00E+00	2.38E-05	4.07E-15

4. Numerical results. Algorithm IDLN was implemented in FORTRAN and run on a DECstation 3100 under the Ultrix V4.0 Operating System. The source code was compiled by using the “-O” option. All floating-point operations were done in double precision. All times reported here were obtained by calling the system subroutine *etime*().

It is not practical to convert the constraints of a linear program to be solved to the form described in Remark 18. We have implemented the algorithm to solve the linear program in standard form (9), where A is a general $m \times n$ matrix. For all test problems, we set the initial values of ε and γ to

$$\varepsilon^0 = 10^{-3}, \quad \gamma^0 = 10^{-6}.$$

The initial value of the dual variable u is

$$u^0 = 0.0.$$

TABLE 5
Comparison of MINOS 5.3 and IDLN (DECstation 3100).

Pr. No.	Problem Name	MINOS 5.3 Obj. Value	IDLN Obj. Value	Rel. Error
1	25fv47	5.501846779100E+03	5.501845888285E+03	1.62E-07
2	Adlittle	2.254949631624E+05	2.254949631627E+05	1.45E-12
3	Afiro	-4.647531428571E+02	-4.647531428543E+02	6.18E-12
4	Agg	-3.599176728658E+07	-3.599176728657E+07	2.87E-13
5	Agg2	-2.023925235598E+07	-2.023925235598E+07	7.20E-14
6	Agg3	1.031211593509E+07	1.031211593922E+07	4.01E-10
7	BandM	-1.586280184501E+02	-1.586280184497E+02	2.43E-12
8	Beaconfd	3.359248580720E+04	3.359248580720E+04	6.93E-15
9	Blend	-3.081214984583E+01	-3.081214984570E+01	4.31E-12
10	Bnl1	1.977629285606E+03	1.977629571770E+03	1.45E-07
11	Bnl2	1.811237723508E+03	1.811236540359E+03	6.53E-07
12	Bore3d	1.373080394208E+03	1.373080394209E+03	7.62E-15
13	BrandY	1.518509896488E+03	1.518509896488E+03	1.08E-14
14	Capri	2.690012913768E+03	2.690012913773E+03	1.97E-12
15	Cre-a	2.359541131854E+07	2.359540706121E+07	1.80E-07
16	Cre-c	2.527511614088E+07	2.527511631673E+07	6.96E-09
17	CzProb	2.185196698857E+06	2.185196698857E+06	2.50E-13
18	D2q06c	1.227842228276E+05	1.227842108308E+05	9.77E-08
19	Degen2	-1.435178000000E+03	-1.435177999966E+03	2.35E-11
20	Degen3	-9.872940000000E+02	-9.872940000029E+02	2.90E-12
21	E226	-1.875192906637E+01	-1.875192906125E+01	2.73E-10
22	Fffff800	5.556795691272E+05	5.556795660507E+05	5.54E-09
23	Finnis	1.727909654670E+05	1.727912302152E+05	1.53E-06
24	Gfrd-Pnc	6.902235999549E+06	6.902235999549E+06	1.51E-14
25	Grow15	-1.068709412936E+08	-1.068709412936E+08	0.00E+00
26	Grow22	-1.608343364826E+08	-1.608343364825E+08	2.87E-13
27	Grow7	-4.778781181471E+07	-4.778781181471E+07	1.56E-16
28	Israel	-8.966448218630E+05	-8.966448157000E+05	6.87E-09
29	Kb2	-1.749900129906E+03	-1.749900129906E+03	3.30E-13
30	Lotfi	-2.526470606188E+01	-2.526470606185E+01	1.10E-12
31	Pilot.we	-2.720104227125E+06	-2.720107532760E+06	1.22E-06
32	Rabo	6.651024202721E+04	6.651024150298E+04	7.88E-09
33	Recipe	-2.666160000000E+02	-2.666160000000E+02	7.68E-15

For most of the problems solved, the initial value of the dual variable v is

$$v^0 = 60.0.$$

These initial values, together with the following updating schemes for the parameters γ and ϵ , produced the best computational results that we could obtain. The parameters are updated as follows. If $\epsilon^i > \epsilon_{\min} = 10^{-12}$, then

$$\epsilon^{i+1} = \epsilon^i / 4.0,$$

and, if $\gamma^i > \gamma_{\min} = 10^{-19}$, then

$$\gamma^{i+1} = \begin{cases} \gamma^i / 1.2 & \text{if } \|x^{i+1} - x^i\|^2 > 100,000, \\ \gamma^i / 2.0 & \text{if } \|x^{i+1} - x^i\|^2 > 1,000, \\ \gamma^i / 3.0 & \text{if } \|x^{i+1} - x^i\|^2 > 10, \\ \gamma^i / 3.5 & \text{if } \|x^{i+1} - x^i\|^2 > 0.1, \\ \gamma^i / 4.0 & \text{otherwise.} \end{cases}$$

TABLE 6
Comparison of MINOS 5.3 and IDLN (DECstation 3100) (continued).

Pr. No.	Problem Name	MINOS 5.3 Obj. Value	IDLN Obj. Value	Rel. Error
34	Sc105	-5.220206121171E+01	-5.220206121090E+01	1.55E-11
35	Sc205	-5.220206121171E+01	-5.220206121170E+01	1.61E-13
36	Sc50a	-6.457507705856E+01	-6.457507705241E+01	9.53E-11
37	Sc50b	-7.000000000000E+01	-6.999999990671E+01	1.33E-09
38	Scagr25	-1.475343306077E+07	-1.475343306075E+07	1.38E-12
39	Scagr7	-2.331389752379E+06	-2.331389822188E+06	2.99E-08
40	Scfxm1	1.841675902835E+04	1.841675902838E+04	1.49E-12
41	Scfxm2	3.666026156500E+04	3.666026156912E+04	1.12E-10
42	Scfxm3	5.490125454975E+04	5.490125454978E+04	4.85E-13
43	Scorpion	1.878124822738E+03	1.878124822709E+03	1.55E-11
44	Scrs8	9.042999861889E+02	9.042969538008E+02	3.35E-06
45	ScSd1	8.666666674333E+00	8.666666695922E+00	2.49E-09
46	ScSd6	5.050000007826E+01	5.050000007863E+01	7.24E-12
47	ScSd8	9.04999999255E+02	9.04999999351E+02	1.07E-11
48	ScTap1	1.412250000000E+03	1.412249999999E+03	4.60E-13
49	ScTap2	1.724807142857E+03	1.724807142857E+03	3.30E-15
50	ScTap3	1.424000000000E+03	1.424000000000E+03	3.83E-14
51	Share1b	-7.658931857919E+04	-7.658931857918E+04	5.83E-14
52	Share2b	-4.157322407414E+02	-4.157322407412E+02	5.11E-13
53	Ship041	1.793324537970E+06	1.793324537970E+06	3.51E-14
54	Ship04s	1.798714700445E+06	1.798714700440E+06	2.82E-12
55	Ship081	1.909055211389E+06	1.909055211390E+06	6.43E-13
56	Ship08s	1.920098210535E+06	1.920098210535E+06	3.12E-13
57	Ship121	1.470187919329E+06	1.470187919330E+06	3.27E-13
58	Ship12s	1.489236134406E+06	1.489236134407E+06	3.16E-13
59	Stocfor1	-4.113197621944E+04	-4.113197620844E+04	2.67E-10
60	Stocfor2	-3.902440853788E+04	-3.902440853788E+04	3.92E-14
61	Truss1	1.143641882582E+04	1.143641313033E+04	4.98E-07
62	Truss2	7.275236330356E+04	7.275236330358E+04	2.66E-13
63	Truss3	4.588158471856E+05	4.588158473024E+05	2.54E-10
64	Vtp. base	1.298314624614E+05	1.298314624614E+05	5.49E-15
65	Wood1p	1.442902411573E+00	1.442902423830E+00	8.49E-09
66	Woodw	1.304476333084E+00	1.304476786146E+00	3.47E-07

The new iterate (u^{i+1}, v^{i+1}) is obtained as follows:

$$\begin{aligned} u^{i+1} &:= \bar{u}^i, \\ v^{i+1} &:= v^i + 0.98\lambda(\bar{v}^i - v^i), \end{aligned}$$

where (\bar{u}^i, \bar{v}^i) is the solution of the linear system (16) and the step size λ is defined as

$$(58) \quad \lambda := \begin{cases} 1 & \text{if } \bar{v}^i \geq 0, \\ \min_{j \in J} (v_j^i / (v_j^i - \bar{v}_j^i)) & \text{otherwise,} \end{cases}$$

where $J := \{j \mid v_j^i - \bar{v}_j^i > 0\}$.

Finally, the program is terminated if one of the following conditions is satisfied:

$$|(cx^i - cx^{i-1}) / cx^{i-1}| \leq 5 \times 10^{-8} \quad \text{and} \quad \|(-x)_+\| / \|x_+\| \leq 10^{-7}$$

or

$$|(cx^i - cx^{i-1}) / cx^{i-1}| \leq 10^{-8} \quad \text{and} \quad |(bu^i - bu^{i-1}) / bu^{i-1}| \leq 10^{-8}.$$

TABLE 7
Time comparison of MINOS 5.3 and IDLN (DECstation 3100).

Pr. No.	Problem Name	IDLN Iter.	MINOS 5.3 (seconds)	IDLN (seconds)	MINOS/IDLN Time Ratio
1	25fv47	63 (5)	329.27	145.02	2.27
2	Adlittle	26 (1)	0.99	1.03	0.96
3	Afiro	23 (1)	0.39	0.66	0.59
4	Agg	40 (3)	4.38	26.08	0.17
5	Agg2	31 (2)	7.40	31.96	0.23
6	Agg3	32 (2)	7.46	33.03	0.23
7	BandM	38 (2)	10.38	7.55	1.37
8	Beaconfd	30 (2)	3.27	6.84	0.48
9	Blend	31 (1)	1.13	1.57	0.72
10	Bnl1	47 (2)	41.31	26.15	1.58
11	Bnl2	68 (5)	608.77	872.00	0.70
12	Bore3d	36 (1)	2.95	4.87	0.61
13	BrandY	45 (2)	6.27	7.57	0.83
14	Capri	48 (1)	4.52	13.28	0.34
15	Cre-a	63 (2)	580.86	139.35	4.17
16	Cre-c	71 (3)	648.83	132.53	4.90
17	Czprob	60 (1)	72.27	34.74	2.08
18	D2q06c	54 (3)	6159.31	1240.51	4.97
19	Degen2	31 (2)	25.55	28.26	0.90
20	Degen3	39 (2)	684.18	665.93	1.03
21	E226	53 (2)	7.34	9.31	0.79
22	Ffff800	51 (3)	25.90	50.77	0.51
23	Finnis	40 (3)	10.73	11.96	0.90
24	Gfrd-Pnc	35 (2)	18.66	7.31	2.55
25	Grow15	32 (2)	17.90	18.73	0.96
26	Grow22	31 (2)	33.46	28.53	1.17
27	Grow7	30 (2)	4.81	8.04	0.60
28	Israel	47 (3)	4.00	37.51	0.11
29	Kb2	32 (2)	0.63	1.15	0.55
30	Lotfi	35 (2)	3.63	2.90	1.25
31	Pilot.we	93 (3)	220.98	98.76	2.24
32	Rabo	65 (2)	15.72	126.82	0.12
33	Recipe	44 (1)	1.10	2.49	0.44

At the termination of Algorithm IDLN, an iterative scheme described in [9] is implemented to improve the accuracy of the solution of the linear program.

We tested the algorithm on 66 linear test problems, 63 of which were from the NETLIB collection [5], [8]. Problem Rabo came from the mortgage division of Rabo Bank of the Netherlands, and the problems Cre-a and Cre-c were made available by J. Kennington [2]. The dimensions of these 66 problems are given in Tables 1 and 2. In columns 3–5 of these tables, we list the number of rows (including the objective row), columns, and nonzeros of matrix A of the linear program in its original MPS format. Columns 6–8 show the size of the linear programs after the data is preprocessed so that these linear programs can be written in standard format (9).

For comparison, we solved these problems by using MINOS 5.3 [24], which is a linear-programming package based on the simplex method. MINOS was run with the default settings for all parameters except *log frequency* 200, *summary frequency* 200, and *solution no.* The results that we obtained on the 66 test problems are listed in Tables 3–8.

TABLE 8
Time comparison of MINOS 5.3 and IDLN (DECstation 3100) (continued).

Pr. No.	Problem Name	IDLN Iter.	MINOS 5.3 (seconds)	IDLN (seconds)	MINOS/IDLN Time Ratio
34	Sc105	30 (2)	0.89	1.10	0.81
35	Sc205	34 (2)	2.02	1.90	1.06
36	Sc50a	27 (2)	0.46	0.77	0.60
37	Sc50b	25 (1)	0.44	0.75	0.59
38	Scagr25	34 (2)	8.20	4.68	1.75
39	Scagr7	32 (2)	1.26	1.40	0.90
40	Scfxm1	42 (1)	7.63	8.33	0.92
41	Scfxm2	45 (2)	21.71	19.10	1.14
42	Scfxm3	47 (2)	44.64	31.10	1.44
43	Scorpion	32 (1)	4.54	3.65	1.24
44	Scrs8	51 (2)	18.68	12.89	1.45
45	ScSd1	27 (1)	4.44	3.17	1.40
46	ScSd6	33 (1)	17.68	6.48	2.73
47	ScSd8	29 (1)	92.62	12.52	7.40
48	ScTap1	39 (2)	4.56	4.62	0.99
49	ScTap2	40 (1)	29.26	31.40	0.93
50	ScTap3	40 (2)	59.14	40.68	1.45
51	Sharelb	45 (2)	2.81	2.81	1.00
52	Share2b	31 (2)	1.59	1.77	0.90
53	Ship041	29 (2)	11.61	10.05	1.16
54	Ship04s	30 (1)	7.32	7.23	1.01
55	Ship081	32 (3)	29.65	23.71	1.25
56	Ship08s	31 (2)	16.71	13.14	1.27
57	Ship12l	32 (1)	66.32	32.31	2.05
58	Ship12s	32 (2)	28.78	16.48	1.75
59	Stocfor1	26 (2)	1.07	1.38	0.78
60	Stocfor2	37 (2)	178.07	45.26	3.93
61	Truss1	34 (1)	23.09	12.09	1.91
62	Truss2	37 (2)	167.08	68.80	2.43
63	Truss3	45 (5)	902.58	212.97	4.24
64	Vtp.base	30 (1)	2.30	4.28	0.54
65	Wood1p	61 (4)	116.13	862.77	0.13
66	Woodw	65 (5)	316.53	265.82	1.19
—	TOTAL	—	11754.16	5588.62	2.10

In Tables 3 and 4, we list

$$\text{Primal Infeasibility} = \max\left(\frac{\|Ax - b\|}{\|b\|}, \|(-x)_+\|\right),$$

$$\text{Dual Infeasibility} = \frac{\|(A'u - c)_+\|}{\|(-c)_+\| + 1.0},$$

$$\text{Duality Gap} = \left| \frac{cx - bu}{cx + bu} \right|,$$

$$\text{Complementarity} = \frac{\|X(c - A'u)\|}{\|x\| \|u\|},$$

where $X := \text{diag}(x)$.

Tables 5 and 6 show the objective values obtained by MINOS and IDLN. Column 5 of these tables gives the relative accuracy of IDLN objective value

$$\text{Relative Error} := \left| \frac{cx - cx^*}{cx^*} \right|,$$

where cx^* is the optimal objective value reported by MINOS.

Tables 7 and 8 compare the execution times for MINOS and IDLN. The total number of iterations for IDLN to solve each problem is shown in column 3, where the number in parentheses indicates the number of refinement iterations.

We note that, for most problems, IDLN solutions have better primal feasibility than do the solutions obtained by the IPP algorithm described in [27]. In the primal algorithm, a Newton direction is computed in the primal space, i.e., the descent direction p is such that $Ap = 0$ and the primal variable is updated $x^{i+1} = x^i + \alpha p$. As i increases, the error $\|Ax^i - b\|$ accumulates, and this leads to a deterioration in the feasibility of the primal solution. In contrast, because Algorithm IDLN takes the Newton step in the dual space, its primal feasibility $Ax^i = b$ depends only on the accuracy of the current Newton direction.

The results obtained from these 66 linear programs can be summarized as follows. After relatively few refinement iterations (< 3 for most problems), solutions with very good primal and dual feasibilities were obtained. For most problems solved, the IDLN objective value agreed with the MINOS objective value at least in the first eight digits. IDLN solved 34 of the 66 problems faster than did MINOS. Similar to other interior-point algorithms, the relative speedup of IDLN over MINOS increases as the problem dimension grows. The total time taken by IDLN to solve all the test problems was 5,589 seconds, whereas the total time for MINOS 5.3 to solve these problems was 11,754 seconds, which is 2.10 times as long as for IDLN.

The number of iterations for Algorithm IDLN is slightly more than that for the two-phase affine scaling algorithm of Adler, Resende, and Veiga [1], but it is very comparable to the number for the single-phase dual-barrier method of Gill, Murray, and Saunders [11]. To solve 31 NETLIB test problems, a total of 1005 iterations, 271 of which are Phase I iterations, are needed by the affine scaling algorithm. The total number of iterations for IDLN to solve these 31 problems is 1172. The corresponding total number of iterations for the single-phase dual-barrier method is 1097.

5. Summary. We have described a new algorithm for finding the least 2-norm solution of a linear program. The logarithmic-penalty approach is applied to the dual reformulation of the problem to find this solution. The dual problem has only nonnegativity constraints on some of the variables; hence finding an initial point for the algorithm and maintaining positivity of these variables is trivial. When the algorithm is started with an appropriate initial point and parameters, and if the parameters are updated at each iteration in a certain prescribed way, the algorithm is shown to converge globally and linearly locally.

Our numerical results indicate that the algorithm is competitive with other interior-point algorithms, such as the affine algorithm [1], [23] and the single-phase dual-barrier method [11].

REFERENCES

- [1] I. ADLER, M. G. C. RESENDE, AND G. VEIGA (1989), *An implementation of Karmarkar's algorithm for linear programming*, Math. Programming, 44, pp. 297-335.
- [2] W. J. CAROLAN, J. E. HILL, J. L. KENNINGTON, S. NIEMI, AND S. J. WICHMAN (1989), *An empirical evaluation of the KORB algorithms for the military airlift applications*, Tech. Report 89-OR-06, Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX.
- [3] R. DE LEONE AND O. L. MANGASARIAN (1987), *Serial and parallel solution of large scale linear programs by augmented Lagrangian successive overrelaxation*, in Optimization, Parallel Processing and Applications 304, A. Kurzhanski, K. Neumann, and D. Pallaschke, eds., Springer-Verlag, Berlin, New York.
- [4] J. E. DENNIS, JR., A. M. MORSHEDI, AND K. TURNER (1987), *A variable-metric variant of the Karmarkar algorithm for linear programming*, Math. Programming, 39, pp. 1-20.
- [5] J. J. DONGARRA AND E. GROSSE (1985), *Distribution of mathematical software via electronic mail*, SIGNUM Newsletter, 20, pp. 45-47.
- [6] S. C. EISENSTAT, M. C. GURSKY, M. H. SCHULTZ, AND A. H. SHERMAN (1982), *Yale Sparse Matrix Package I: The symmetric codes*, Internat. J. Numer. Meth. Engrg., 18, pp. 1145-1151.
- [7] ——— (1977), *Yale Sparse Matrix Package I: The symmetric codes*, Research Report 112, Yale University, New Haven, CT.
- [8] D. M. GAY (1985), *Electronic main distribution of linear programming test problems*, Math. Programming Soc. COAL Newsletter, December.
- [9] ——— (1989), *Stopping tests that computer optimal solutions for interior-point linear programming algorithms*, manuscript, AT&T Bell Laboratories, Murray Hill, NJ.
- [10] P. E. GILL, W. MURRAY, M. A. SAUNDERS, J. A. TOMLIN, AND M. H. WRIGHT (1986), *On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method*, Math. Programming, 36, pp. 183-209.
- [11] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS (1988), *A single-phase dual barrier method for linear programming*, Tech. Report SOL 88-10, Department of Operations Research, Stanford University, Stanford, CA.
- [12] M. KOJIMA, S. MIZUNO, AND A. YOSHISE (1989), *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Programming, 44, pp. 1-26.
- [13] I. J. LUSTIG (1988), *A generic primal-dual interior point algorithm*, Tech. Report SOR 88-3, Department of Civil Engineering and Operations Research, Princeton University, Princeton, NJ.
- [14] O. L. MANGASARIAN (1969), *Nonlinear Programming*, McGraw-Hill, New York.
- [15] ——— (1977), *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22, pp. 465-485.
- [16] O. L. MANGASARIAN AND R. R. MEYER (1979), *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17, pp. 745-757.
- [17] O. L. MANGASARIAN (1981), *A stable theorem of the alternative: An extension of the Gordan theorem*, Linear Algebra Appl., 41, pp. 209-223.
- [18] ——— (1984), *Normal solutions of linear programs*, Math. Programming Study, 22, pp. 206-216.
- [19] O. L. MANGASARIAN AND R. DE LEONE (1986), *Parallel gradient projection successive overrelaxation for symmetric linear complementarity problems and linear programs*, Tech. Report 659, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI.
- [20] O. L. MANGASARIAN AND T.-H. SHIAU (1987), *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Control Optim., 25, pp. 583-595.
- [21] O. L. MANGASARIAN AND R. DE LEONE (1988), *Error bounds for strongly convex programs and (super) linearly convergent iterative schemes for the least 2-norm solution of linear programs*, Appl. Math. Optim., 17, pp. 1-14.
- [22] K. A. MCSHANE, C. L. MONMA, AND D. SHANNO (1989), *An implementation of primal-dual interior point method for linear programming*, ORSA J. Comput., 1, pp. 70-83.
- [23] C. L. MONMA AND A. J. MORTON (1987), *Computational experience with a dual affine variant of the Karmarkar's method for linear programming*, Oper. Res. Lett., 6, pp. 261-267.
- [24] B. A. MURTAGH AND M. A. SAUNDERS (1983), *MINOS 5.0 user's guide*, Tech. Report SOL 83-20, Stanford Optimization Laboratory, Stanford, CA.
- [25] J. M. ORTEGA AND W. C. RHEINOLDT (1970), *Iterative Solution of Nonlinear Complementarity Equations in Several Variables*, Academic Press, New York.
- [26] B. T. POLYAK (1987), *Introduction to Optimization*, Optimization Software, New York, 1987.
- [27] R. SETIONO (1990), *Interior proximal point algorithm for linear programs*, Tech. Report 949, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI.

- [28] P. TSENG (1988), *A simple polynomial-time algorithm for convex quadratic programming*, Report LIDS-P-1819, Center for Intelligent Control Systems, Massachusetts Institute of Technology, Cambridge, MA.
- [29] ——— (1990), *A path-following algorithm for linear programming using quadratic and logarithmic penalty functions*, Report LIDS-P-1963, Center for Intelligent Control Systems, Massachusetts Institute of Technology, Cambridge, MA.

CONTROLLABILITY ALONG A TRAJECTORY: A VARIATIONAL APPROACH*

ROSA MARIA BIANCHINI† AND GIANNA STEFANI‡

Abstract. This paper unifies and improves most sufficient conditions of local controllability both along a trajectory and at a point. This is accomplished by defining high-order variations that can be continuously summed. The peculiar property of these variations is that they may be generated by thin conditions as relations in the Lie algebra associated with a control system. This property leads to a variational interpretation of Sussmann-type sufficient conditions of local controllability (neutralization of obstructions), and it allows the use of different weights for neutralization.

Key words. local controllability at a point, local controllability along a trajectory, high-order variations, C^1 control systems, C^∞ affine control systems, graded approximations

AMS subject classifications. 49E15, 93B05, 93C10, 93C45

Introduction. The aim of this paper is to give high-order conditions for a point of a reference trajectory to be interior to the reachable set. This property is linked to the minimum-time problem by the fact that if \hat{x} is an optimal trajectory on $[0, T]$, then, for each $t \in [0, T)$, $\hat{x}(t)$ belongs to the boundary of $R(\xi_0, t)$, i.e., the reachable set at time t from the initial point $\xi_0 = \hat{x}(0)$ of the optimal trajectory. Therefore, each sufficient condition for $\hat{x}(t)$ to belong to the interior of $R(\xi_0, t)$ yields a necessary condition for \hat{x} to be time optimal. The high-order conditions are of particular interest in the case of nonlinear systems for which the Pontryagin maximum principle may not be sufficient to single out a unique candidate and singular trajectories may appear.

The original motivation for the paper was to give a unified setting to most of the results on local controllability, both at a point and along a reference trajectory. For the controllability at a point ξ_0 of systems with bounded controls, a quite general sufficient condition was obtained by the same authors in [1] by means of an open mapping theorem based on the properties of the cone of tangent directions to the reachable set at ξ_0 given in [7]. For controllability along a possibly nonstationary trajectory, the idea is to use some known conditions based on the relations at a point in the Lie algebra associated with the system in order to construct high-order variations of the trajectory.

Variational cones \mathcal{H} have been introduced to state high-order maximum principles [5], [8], [13], [14], etc., for optimal-control problems in a fixed interval of time $[0, T]$. If there are constraints on the final point, then it is usually required that the elements of \mathcal{H} have a continuous sum property (see [5], [14]). Roughly speaking, the property can be stated as follows: if v_1, \dots, v_k belong to \mathcal{H} , then for all sufficiently small positive numbers c_1, \dots, c_k there are control variations depending continuously on the data that produce a curve in the reachable set $R(\xi_0, T)$ whose tangent vector is given by $c_1 v_1 + c_2 v_2 + \dots + c_k v_k$. The continuous-sum property ensures that if \mathcal{H} is the whole tangent space, then the final point $\hat{x}(T)$ is interior to $R(\xi_0, T)$.

Control variations, possibly distributed along the trajectory, are defined in [5]. The result is that a variation of high order can be summed only with variations of order one. In [13], [14], instantaneous control variations are considered, i.e., the control

* Received by the editors September 4, 1990; accepted for publication (in revised form) February 25, 1992.

† Dipart. di Matematica U. Dini, Viale Morgagni 67/a, 50134 Firenze, Italy.

‡ Dipart. di Matematica e Appl., Via Mezzocannone 8, 80100 Napoli, Italy.

variation takes place in a small interval of time $[\tau, \tau + \varepsilon]$. For this kind of variation the main difficulties in proving the continuous sum property arise for tangent vectors generated by control variations that occur at the same time τ . In [13] and [14] these difficulties are overcome by requiring that a variation can be obtained continuously on a small arc of trajectory, so that the variations can be thought of as being produced at different times. This implies that the variations are not of local nature, so that they cannot be produced by thin conditions, such as those based on the relations that the Lie algebra associated with the system satisfies at a point on the trajectory; see Example 3.10. In [8] a high-order maximum principle is stated for unconstrained optimization problems. The variations defined therein have a local nature, but they do not have the continuous-sum property.

In this paper we define variations of a reference trajectory that satisfy the continuous-sum property and have a local nature. This is achieved by requiring that control variations of time duration ε and starting at times $\tau + \gamma\varepsilon$ (γ is a new “small” parameter) produce basically the same tangent vector to the reachable sets. This variational approach improves and unifies most of the known sufficient conditions for both local controllability at a point and local controllability along a reference trajectory [1], [3], [9], [12], [17], [18]. In particular, we prove that the Sussmann-type controllability conditions give rise to variations. As a consequence, we use different weights in neutralizing different obstructions. Our techniques can be used for both bounded and unbounded controls. Therefore, the results in [1] are also improved; in fact, the open mapping theorem described therein applies only to bounded controls. For a preliminary version of the results contained in this paper, see [2]. The paper is organized as follows: In § 1 we define the variational cone and state its main properties. In § 2 we study further properties of the variations for C^∞ systems, and we introduce the notion of weak local controllability. In §§ 3, 4, and 5, C^∞ control systems that are affine with respect to the control are considered. More precisely, in § 3 we prove that the variations depend on the “germ” of the system (Proposition 3.1), and we state the main result (Theorem 3.5) concerning the possibility of producing variations by means of the relations in the Lie algebra associated with the system. Many examples are given. In § 4 we study the variations of the trajectory relative to the drift term by using an approximating system we defined in [3]. In § 5 we give the applications to local controllability along a reference trajectory, Theorem 5.4. This theorem points out that controllability at a point and controllability along a trajectory can be seen in a unified setting.

1. The variational cone. Let M be an n -dimensional, paracompact connected C^q manifold, $q \geq 2$. We use the following notation: If $h: M \rightarrow M$ is a diffeomorphism, $h_*: TM \rightarrow TM$ is the tangent map of h . $o(\cdot)$ denotes a continuous map such that $\lim_{\varepsilon \rightarrow 0} o(\varepsilon)/\varepsilon = 0$, and $O(\cdot)$ denotes a continuous map such that $\lim_{\varepsilon \rightarrow 0} O(\varepsilon) = 0$. Let $\eta: [0, \bar{\varepsilon}] \rightarrow M$ be a continuous map such that $\eta(0) = \xi$, and let $v \in T_\xi M$. We write

$$\eta(\varepsilon) = \xi + \varepsilon^k v + o(\varepsilon^k)$$

if the equality holds in a chart (and hence in any chart). A time-dependent vector field $g: \mathbb{R} \times M \rightarrow TM$ is said to be a C^r Caratheodory vector field [10] if $q \geq r + 1$ and the following conditions are fulfilled:

(a) $\forall t \in \mathbb{R}$, $g(t, \cdot)$ is C^r .

(b) In any chart, g and all its derivatives with respect to $x \in M$ up to order r are measurable in t . To be more precise, if $D_x^j g$ denotes the j th derivative of g with respect to $x \in M$, then $D_x^j g(\cdot, x)$ is measurable $\forall x \in M$, $j = 0, 1, \dots, r$.

(c) In each chart, $D_2^j g, j = 0, 1, \dots, r$, is locally L^1 -bounded, i.e., for each compact K contained in the chart and for each $j = 1, \dots, r$ there exists a locally integrable map ϕ_j such that

$$\|D_2^j g(t, x)\| \leq \phi_j(t).$$

Let us consider a multi-input control system on the manifold M , i.e., a differential equation Σ on M :

$$\dot{x} = f(t, x, u),$$

depending on the control map $u \in \mathcal{U}$. \mathcal{U} will be called the set of admissible controls. We assume the following:

(A1) Each control map u is a measurable map defined on a compact interval I_u with values in \mathbb{R}^m .

(A2) $f: \mathbb{R} \times M \times \mathbb{R}^m \rightarrow TM$ is such that for each fixed $(t, \omega) \in \mathbb{R} \times \mathbb{R}^m$, $f(t, \cdot, \omega)$ is a C^1 map and for each $u \in \mathcal{U}$, $(t, x) \mapsto f(t, x, u(t))$ is a C^1 Caratheodory vector field.

(A3) If u belongs to \mathcal{U} , then each restriction of u to a subinterval of I_u belongs to \mathcal{U} . We shall denote both the control and its restriction by the same symbol.

(A4) If u, v belong to \mathcal{U} and $I_u = [a, b]$, $I_v = [b, c]$, then the concatenation $u \# v: [a, c] \rightarrow \mathbb{R}^m$ defined by

$$u \# v(t) = \begin{cases} u(t), & t \in [a, b], \\ v(t), & t \in [b, c], \end{cases}$$

belongs to \mathcal{U} .

Under these assumptions for each $u \in \mathcal{U}$, each $\xi \in M$, and each $t_0 \in I_u$, there is a unique maximal solution $t \mapsto S(t, t_0, \xi, u)$ of Σ such that $S(t_0, t_0, \xi, u) = \xi$. The map S defined on a suitable subset of $\mathbb{R} \times \mathbb{R} \times M \times \mathcal{U}$ by $(t, t_0, \xi, u) \mapsto S(t, t_0, \xi, u)$ is called the flow of the control system. For each fixed u , S is a continuous map that is C^1 with respect to ξ . We endow \mathcal{U} with a topology Π for which the flow and its derivative with respect to ξ , $(t, t_0, \xi, u) \mapsto (S(t, t_0, \cdot, u))_*$ are continuous. The possible choices of the topology Π depend on f and on the properties of the admissible controls [5], [10], [11]. If \mathcal{U} is immersed in $L^\infty(\mathbb{R})$, then we can choose Π equal to the L^∞ topology [10]. If f is affine in u , then the L^1 topology can be chosen as Π also; see [5].

Let us fix a reference control \hat{u} , and without loss of generality let us suppose $I_{\hat{u}} = [0, T]$. The reference control \hat{u} defines a reference flow

$$(t, \tau, \xi) \mapsto S(t, \tau, \xi, \hat{u}) \equiv \hat{S}(t, \tau, \xi).$$

Let us fix also the initial point ξ_0 , so that the reference trajectory $t \mapsto \hat{x}(t) \equiv \hat{S}(t, 0, \xi_0)$ is completely determined.

Definition 1.1. The control system is said to be locally controllable along \hat{x} (or, if the system can be understood, \hat{x} is locally controllable) if and only if for each $t \in (0, T]$

$$\hat{x}(t) \in \text{int } R(\xi_0, t) = \text{int}\{S(t, 0, \xi_0, u): u \in \mathcal{U}\}.$$

In this section we define the variational cone relative to the reference couple (\hat{x}, \hat{u}) , and we state its main properties. As a consequence, we will get conditions for the point $\hat{x}(t)$ to be interior to the reachable set and hence for the reference trajectory to be locally controllable. Since the proofs of the theorems are rather technical, we provide them at the end of the section.

We start by defining a variation of a reference trajectory at a given time t . Roughly speaking, we consider the principal part with respect to ε of the curve obtained by first applying a control variation starting at t for an interval of time length ε and then

transporting the obtained final point back to time t by means of the reference flow. Therefore, a control variation results in a tangent vector at $\hat{x}(t)$. We require that the same principal part can be obtained if the starting point of the control variation is shifted by $\varepsilon\gamma$, where γ is a small parameter that is independent of ε . It is this last property that allows us to sum the simultaneous variations (see the proofs of Lemma 1.15 and Theorem 1.7).

DEFINITION 1.2. A vector $v \in T_{\hat{x}(t)}M$ is a *right variation* of order α of (\hat{x}, \hat{u}) at time $t \in [0, T)$ if there are (a) positive numbers $\bar{\gamma}, \bar{c}, \bar{\varepsilon}$ and (b) a map $\eta: [0, \bar{\gamma}] \times [0, \bar{c}] \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$ continuous in the Π topology of \mathcal{U} such that

$$(1.1) \quad \hat{S}(t, t + \varepsilon\gamma + \varepsilon, S(t + \varepsilon\gamma + \varepsilon, t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \eta(\gamma, c, \varepsilon))) = \hat{x}(t) + \varepsilon^\alpha cv + o(\varepsilon^\alpha)$$

uniformly with respect to γ and c . The vector v is a *left variation* at time $t \in (0, T]$ if

$$(1.2) \quad \hat{S}(t, t - \varepsilon\gamma, S(t - \varepsilon\gamma, t - \varepsilon\gamma - \varepsilon, \hat{x}(t - \varepsilon\gamma - \varepsilon), \eta(\gamma, c, \varepsilon))) = \hat{x}(t) + \varepsilon^\alpha cv + o(\varepsilon^\alpha).$$

The variation will be called *regular* if the map $o(\varepsilon^\alpha)$ that occurs in (1.1) or (1.2) is an $o(\varepsilon^\beta)$ for some $\beta > \alpha$.

Example 1.3. If the constant map $u(t) \equiv \omega$ is an admissible control, then

$$f(t, \hat{x}(t), \omega) - f(t, \hat{x}(t), \hat{u}(t))$$

is a right (left) variation of (\hat{x}, \hat{u}) at each t in which \hat{u} is right (left) continuous.

Remark 1.4. In a chart at $\hat{x}(t)$ equation (1.1) can also be written

$$S(t + \varepsilon\gamma + \varepsilon, t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \eta(\gamma, c, \varepsilon)) = \hat{x}(t + \varepsilon\gamma + \varepsilon) + \varepsilon^\alpha cv + o(\varepsilon^\alpha).$$

Remark 1.5. There are no left variations at $t = 0$ and no right variations at $t = T$.

Remark 1.6. A variation depends only on the reference trajectory and does not depend on the reference control. Namely, if the controls \hat{u} and $\hat{\hat{u}}$ give rise to the same trajectory \hat{x} in a neighborhood of t , then the set of variations at t is the same if either \hat{u} or $\hat{\hat{u}}$ is used.

By definition, the set of variations at t is a cone with vertex at the origin that is not in general a convex set. However, the subset of variations of the same order is a convex set (see Lemma 1.15). Moreover, the simultaneous variations can be summed in the following sense:

THEOREM 1.7. If v_1, \dots, v_r are variations of (\hat{x}, \hat{u}) at time t , then there exist

- (i) a positive number $\bar{\varepsilon}$,
- (ii) a neighborhood V of 0 in $(\mathbb{R}^+)^r$,
- (iii) a continuous map $\nu: V \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$, and
- (iv) two continuous maps $\Theta, \Phi: [0, \bar{\varepsilon}] \rightarrow \mathbb{R}^+$ that go to zero with ε such that

$$\hat{S}(t, t + \Theta(\varepsilon), S(t + \Theta(\varepsilon), t - \Phi(\varepsilon), \hat{x}(t - \Phi(\varepsilon)), \nu(c, \varepsilon))) = \hat{x}(t) + \varepsilon \sum_{i=1}^r c_i v_i + o(\varepsilon).$$

Notice that Theorem 1.7 implies that the transport along the reference flow of the convex hull of the set of variations at t defines a set of directions tangent to the reachable set. Hence the variations that occur at any time $\tau \in [0, t]$ provide information on the geometry of $R(\xi_0, t)$ at $\hat{x}(t)$. For this reason we define the variational cone at a time t as the convex hull of the transport of the variations obtained at each previous time. Contrary to the variations, the variational cone depends on the choice of the reference control except for $t = 0$.

Definition 1.8. Let $t \in [0, T]$. The variational cone $\mathcal{K}(t)$ of (\hat{x}, \hat{u}) at time t is given by

$$\mathcal{K}(t) = \begin{cases} \text{convex hull } \{v : v \text{ is a right variation at } 0\} & \text{if } t = 0. \\ \text{convex hull } \left\{ \bigcup_{\tau \in [0, t]} (\hat{S}(t, \tau, \cdot))_* v : v \text{ is a variation at } \tau \right\} & \text{if } t \neq 0. \end{cases}$$

Here only right (left) variations have to be considered at 0 (at t).

The vectors of $\mathcal{K}(t)$, $t > 0$, define directions tangent to $R(\xi_0, t)$ at $\hat{x}(t)$. In fact, the following property holds.

THEOREM 1.9. Let $w_1, \dots, w_s \in \mathcal{K}(t)$, $t > 0$; there is a neighborhood V of 0 in $(\mathbb{R}^+)^s$ and a continuous map $\eta : V \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$ such that

$$(1.3) \quad S(t, 0, \xi_0, \eta(c, \varepsilon)) = \hat{x}(t) + \varepsilon \sum_{i=1}^s c_i w_i + o(\varepsilon).$$

The above theorem and degree theory provide in an obvious way the following sufficient condition for $\hat{x}(t)$ to be interior to the reachable set at time t .

THEOREM 1.10. If for some $t > 0$, $\mathcal{K}(t) = T_{\hat{x}(t)}M$, then $\hat{x}(t) \in \text{int } R(\xi_0, t)$.

By definition, it follows that if $\mathcal{K}(0) = T_{\xi_0}M$, then $\mathcal{K}(t) = T_{\hat{x}(t)}M$, $\forall t \in [0, T]$. Therefore, Theorem 1.10 implies the following:

COROLLARY 1.11. If $\mathcal{K}(0) = T_{\xi_0}M$, then the system is locally controllable along \hat{x} .

Example 1.12. Let

$$\dot{x} = Ax + bu, \quad \mathcal{U} = L^\infty(\mathbb{R}, [0, 1])$$

be a linear control system in \mathbb{R}^2 such that A has complex nonreal eigenvalues and $\text{rank}(b, Ab) = 2$. Let $\hat{u} \equiv 0$ be the reference control, and let $\hat{x} : t \mapsto e^{tA}\xi_0$ be the associated reference trajectory. b is a variation at each $s \in [0, t]$, and $(e^{A(t-s)}b)$ is the transport of b from time s up to time t . Therefore, the variational cone $\mathcal{K}(t)$ contains $e^{A\tau}b$ for all τ in $[0, t]$. Let $\alpha \pm i\beta$ be the eigenvalues of A ; if $t > \pi/\beta$, then the set $\{e^{A\tau}b : \tau \in [0, t]\}$ contains a positive basis of \mathbb{R}^2 . Hence $\hat{x}(t)$ is an interior point of $R(\xi_0, t)$ for each t greater than π/β .

The same ideas can be used to compute the time interval in which a single-input, linear, autonomous control system in \mathbb{R}^n is globally controllable with positive controls, if all the eigenvalues of A are simple and complex.

Notice that if $\text{int } R(\xi_0, t)$ is empty, Theorem 1.10 provides no information on $\hat{x}(t)$. Nevertheless, for a large class of control systems $R(\xi_0, t)$ is contained in a possibly lower-dimensional submanifold $N(\xi_0, t)$ and the interior of $R(\xi_0, t)$ relative to $N(\xi_0, t)$ is not empty [4]. In these cases the variational cone can be used to test whether $\hat{x}(t)$ is a relatively interior point (see § 2).

THEOREM 1.13. If $R(\xi_0, t)$ is contained in a submanifold N and $\mathcal{K}(t) = T_{\hat{x}(t)}N$, then $\hat{x}(t)$ is interior to $R(\xi_0, t)$ with respect to N .

Proof. The proof can be obtained easily by noting that if $R(\xi_0, t)$ is contained in a submanifold N , the definition of $\mathcal{K}(t)$ implies that $\mathcal{K}(t)$ is contained in the tangent space of N at $\hat{x}(t)$. \square

Remark 1.14. If all the constant maps with values in a subset Ω of \mathbb{R}^m are admissible controls, then Theorem 1.10 can also be put in a variational setting. In fact, it states that if $\hat{x}(t) \in \partial R(\xi_0, t)$, then there exists a nontrivial covector $\bar{\lambda}$ at $\hat{x}(t)$ such that $\langle \bar{\lambda}, w \rangle \leq 0$ for each $w \in \mathcal{K}(t)$. Hence in our hypothesis the definition of variational cone implies that the solution $\lambda : [0, t] \rightarrow T^*M$ of the adjoint equation, which in local

coordinates is given by

$$\dot{\lambda}(s) = -\lambda(s) \frac{\partial}{\partial x} f(s, \hat{x}(s), \hat{u}(s)), \quad \lambda(t) = \bar{\lambda},$$

is such that

$$(1.4) \quad \langle \lambda(s), f(s, \hat{x}(s), \omega) \rangle \leq \langle \lambda(s), f(s, \hat{x}(s), \hat{u}(s)) \rangle, \quad \forall \omega \in \Omega \text{ a.e. } s \in [0, t],$$

and

$$(1.5) \quad \langle \lambda(s), v \rangle \leq 0, \quad \forall v \in \mathcal{H}(s), \quad \forall s \in [0, t].$$

We end this section with the proofs of Theorems 1.7 and 1.9. To prove Theorem 1.7 we need the following two lemmas.

LEMMA 1.15. *Let v_1, \dots, v_s be right (left) variations of order k of (\hat{x}, \hat{u}) at time t ; there exist (i) positive numbers $\bar{\gamma}, \bar{\varepsilon}, \bar{t}$, (ii) a neighborhood V of 0 in $(\mathbb{R}^+)^s$, and (iii) a continuous map $\mu: [0, \bar{\gamma}] \times V \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$ such that*

$$\begin{aligned} \hat{S}(t, t + \varepsilon\gamma + \varepsilon\bar{t}, S(t + \varepsilon\gamma + \varepsilon\bar{t}, t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \mu(\gamma, c, \varepsilon))) &= \hat{x}(t) + \varepsilon^k \sum_{i=1}^s c_i v_i + o(\varepsilon^k), \\ \left(\hat{S}(t, t - \varepsilon\gamma, S(t - \varepsilon\gamma, t - \varepsilon\gamma - \varepsilon\bar{t}, \hat{x}(t - \varepsilon\gamma - \varepsilon\bar{t}), \mu(\gamma, c, \varepsilon))) &= \hat{x}(t) + \varepsilon^k \sum_{i=1}^s c_i v_i + o(\varepsilon^k) \right). \end{aligned}$$

In other words, the convex set generated by v_1, \dots, v_s is a set of variations of order k .

Proof. We give the proof for right variations; the proof for left variations is analogous. Let us fix a chart at $\hat{x}(t)$ so that we can suppose $M = \mathbb{R}^n$. From Definition 1.2 it follows that there are $\bar{\gamma}_i, \bar{\varepsilon}_i, \bar{c}_i$ and continuous maps $\eta_i: [0, \bar{\gamma}_i] \times [0, \bar{c}_i] \times [0, \bar{\varepsilon}_i] \rightarrow \mathcal{U}$ such that

$$(1.6) \quad S(t + \varepsilon_i \gamma_i + \varepsilon_i, t + \varepsilon_i \gamma_i, \hat{x}(t + \varepsilon_i \gamma_i), \eta_i(\gamma_i, c_i, \varepsilon_i)) = \hat{x}(t + \varepsilon_i \gamma_i + \varepsilon_i) + \varepsilon_i^k c_i v_i + o(\varepsilon_i^k).$$

If $\varepsilon_1, \dots, \varepsilon_s$ are such that

$$(1.7) \quad \varepsilon_1(\bar{\gamma}_1 + 1) \leq \varepsilon_2 \bar{\gamma}_2, \varepsilon_1(\bar{\gamma}_1 + 1) + \varepsilon_2 \leq \varepsilon_3 \bar{\gamma}_3, \dots, \varepsilon_1(\bar{\gamma}_1 + 1) + \varepsilon_2 + \dots + \varepsilon_{s-1} \leq \varepsilon_s \bar{\gamma}_s,$$

then

$$\eta_i \left(\frac{\varepsilon_1(\gamma + 1) + \varepsilon_2 + \dots + \varepsilon_{i-1}}{\varepsilon_i}, c_i, \varepsilon_i \right)$$

is defined for $i = 2, \dots, s$, $\gamma \in [0, \bar{\gamma}_1]$, $c_i \in [0, \bar{c}_i]$, $\varepsilon_i \in [0, \bar{\varepsilon}_i]$. Let H_2, \dots, H_s be the positive numbers defined by

$$\bar{\gamma}_1 + 1 = H_2 \bar{\gamma}_2, H_2(\bar{\gamma}_2 + 1) = H_3 \bar{\gamma}_3, \dots, H_{s-1}(\bar{\gamma}_{s-1} + 1) = H_s \bar{\gamma}_s.$$

If $\varepsilon_1 = \varepsilon$, $\varepsilon_i = \varepsilon H_i$, $i = 2, \dots, s$, then the inequalities (1.7) are satisfied. Let

$$\bar{\gamma} = \bar{\gamma}_1, \quad \gamma_i = \gamma_i(\gamma) = (\gamma + 1 + H_2 + \dots + H_{i-1})/H_i,$$

$$\bar{\varepsilon} = \min \{1, \varepsilon_1, \bar{\varepsilon}_2/H_2, \dots, \bar{\varepsilon}_s/H_s\},$$

$$c = (c_1, \dots, c_s), \quad V_s = [0, \bar{c}_1] \times [0, \bar{c}_2 H_2^k] \times \dots \times [0, \bar{c}_s H_s^k].$$

The map $\mu: [0, \bar{\gamma}] \times V_s \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$ given by

$$\mu(\gamma, c, \varepsilon) = \eta_1(\gamma, c_1, \varepsilon) \# \eta_2(\gamma_2, c_2/H_2^k, \varepsilon H_2) \# \dots \# \eta_s(\gamma_s, c_s/H_s^k, \varepsilon H_s)$$

is continuous. Moreover,

$$\begin{aligned} S(t + \varepsilon(\gamma + 1), t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \mu(\gamma, c, \varepsilon)) \\ = S(t + \varepsilon(\gamma + 1), t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \eta_1(\gamma, c_1, \varepsilon)) \\ = \hat{x}(t + \varepsilon\gamma + \varepsilon) + \varepsilon^k c_1 v_1 + o(\varepsilon^k), \end{aligned}$$

$$\begin{aligned}
& S(t + \varepsilon(\gamma + 1 + H_2), t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \mu(\gamma, \mathbf{c}, \varepsilon)) \\
&= S(t + \varepsilon(\gamma + 1 + H_2), t + \varepsilon(\gamma + 1), \hat{x}(t + \varepsilon(\gamma + 1))) \\
&\quad + \varepsilon^k c_1 v_1 + o(\varepsilon^k), \eta_2(\gamma_2, c_2/H_2^k, \varepsilon H_2)).
\end{aligned}$$

By the definition of Π the derivative of the flow with respect to the initial state is continuous, so that

$$\begin{aligned}
& S(t + \varepsilon(\gamma + 1 + H_2), t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \mu(\gamma, \mathbf{c}, \varepsilon)) \\
&= S(t + \varepsilon(\gamma + 1 + H_2), t + \varepsilon(\gamma + 1), \hat{x}(t + \varepsilon(\gamma + 1)), \eta_2(\gamma_2, c_2/H_2^k, \varepsilon H_2)) \\
&\quad + \left[\frac{\partial}{\partial \xi} S(t, t, \hat{x}(t), \eta_2(\gamma_2, c_2/H_2^k, 0)) + O(\varepsilon) \right] (\varepsilon^k c_1 v_1 + o(\varepsilon^k)) \\
&= \hat{x}(t + \varepsilon(\gamma + 1 + H_2)) + \varepsilon^k (c_1 v_1 + c_2 v_2) + o(\varepsilon^k).
\end{aligned}$$

By similar arguments if $\bar{t} = 1 + H_2 + \dots + H_s = H_s(\gamma_s + 1) - \gamma$, then

$$\begin{aligned}
& S(t + \varepsilon(\gamma + \bar{t}), t + \varepsilon\gamma, \hat{x}(t + \varepsilon\gamma), \mu(\gamma, \mathbf{c}, \varepsilon)) \\
&= S(t + \varepsilon H_s(\gamma_s + 1), t + \varepsilon H_s \gamma_s, \hat{x}(t + \varepsilon H_s \gamma_s) + \varepsilon^k (c_1 v_1 + \dots + c_{s-1} v_{s-1}) \\
&\quad + o(\varepsilon^k), \eta_s(\gamma_s, c_1/H_s^k, \varepsilon H_s)) \\
&= \hat{x}(t + \varepsilon\gamma + \varepsilon \bar{t}) + \varepsilon^k (c_1 v_1 + \dots + c_s v_s) + o(\varepsilon^k),
\end{aligned}$$

so that the lemma is proved. \square

LEMMA 1.16. Let v_1, \dots, v_r be right (left) variations of (\hat{x}, \hat{u}) at time t ; there exist (i) a positive number $\bar{\varepsilon}$, (ii) a neighborhood V of 0 in $(\mathbb{R}^+)^r$, (iii) a continuous map $\nu: V \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$, and (iv) a continuous map $\Theta: [0, \bar{\varepsilon}] \rightarrow \mathbb{R}^+$ that goes to zero with ε such that

$$\begin{aligned}
& \hat{S}(t, t + \Theta(\varepsilon), S(t + \Theta(\varepsilon), t, \hat{x}(t), \nu(c, \varepsilon))) = \hat{x}(t) + \varepsilon \sum_{i=1}^r c_i v_i + o(\varepsilon), \\
& \left(S(t, t - \Theta(\varepsilon), \hat{x}(t - \Theta(\varepsilon)), \nu(c, \varepsilon)) = \hat{x}(t) + \varepsilon \sum_{i=1}^r c_i v_i + o(\varepsilon) \right).
\end{aligned}$$

Proof. Let $k_1 < k_2 < \dots < k_p$ be the orders of the variations v_i 's. We denote by w_{i1}, \dots, w_{is_i} the variation of order k_i . Let $h_i = 1/k_i$; then $h_1 > h_2 > \dots > h_p$. By Lemma 1.15, for each i there are $\tilde{\gamma}_i, \tilde{\varepsilon}_i, \tilde{t}_i, V_i$ and $\mu_i: [0, \tilde{\gamma}_i] \times V_i \times [0, \tilde{\varepsilon}_i] \rightarrow \mathcal{U}$ such that in a chart at $\hat{x}(t)$

$$\begin{aligned}
& S(t + \varepsilon^{h_i} \gamma + \varepsilon^{h_i} \tilde{t}_i, t + \varepsilon^{h_i} \gamma, \hat{x}(t + \varepsilon^{h_i} \gamma), \mu_i(\gamma, d^i, \varepsilon^{h_i})) \\
&= \hat{x}(t + \varepsilon^{h_i} \gamma + \varepsilon^{h_i} \tilde{t}_i) + \varepsilon \sum_{j=1}^{s_i} d_j^i w_{ij} + o(\varepsilon).
\end{aligned}$$

Hence if ε is sufficiently small, say, $\varepsilon \in [0, \varepsilon']$, and if $d^i \in V_i$, then $\nu_i: (\varepsilon, d^i) \mapsto \mu_i(\varepsilon^{h_i} \tilde{t}_1 + \dots + \varepsilon^{h_{i-1}} \tilde{t}_{i-1}, d^i, \varepsilon^{h_i})$ is defined and continuous. Set $\Theta(\varepsilon) = \varepsilon^{h_1} \tilde{t}_1 + \dots + \varepsilon^{h_p} \tilde{t}_p$, $c \in \bigoplus_{i=1}^p V_i$, and $\nu(\varepsilon, c) = \nu_1(\varepsilon, d^1) \# \dots \# \nu_p(\varepsilon, d^p)$. With the same arguments used in the proof of Lemma 1.15 we get

$$S(t + \Theta(\varepsilon), t, \hat{x}(t), \nu(c, \varepsilon)) = \hat{x}(t + \Theta(\varepsilon)) + \varepsilon \sum_{j=1}^r c_j v_j + o(\varepsilon).$$

The proof is analogous for the left variations. \square

Proof of Theorem 1.7. Let v_1, \dots, v_r be variations of (\hat{x}, \hat{u}) at time t , and let us suppose that v_j is a left variation for $j \leq s$ and that it is a right variation for $j > s$. The control variation defined in Lemma 1.16 is a variation of the reference control \hat{u} before the time t for left variations and after time t for right variations. Therefore, by concatenating the two control variations we get the desired property. \square

Proof of Theorem 1.9. Let $\{w_1, \dots, w_s\} \subset \mathcal{H}(t)$; there exists $t_0 < t_1 < \dots < t_p$, $v_{1i}, \dots, v_{ri} \in T_{\hat{x}(t_i)}M$ such that v_{ji} is a variation at time t_i and $\{(S(t, \tau, \cdot, \hat{u}))_* v_{ji}, i = 1, \dots, p, j = 1, \dots, r_i\} = \{w_1, \dots, w_s\}$. For each i let $\bar{\varepsilon}_i, V_i, \Theta_i, \Phi_i$, and ν_i be as in Theorem 1.7; hence in a chart at $\hat{x}(t_i)$

$$(1.8) \quad \begin{aligned} & S(t_i + \Theta_i(\varepsilon), t_i - \Phi_i(\varepsilon), \hat{x}(t_i - \Phi_i(\varepsilon)), \nu_i(c^i, \varepsilon)) \\ &= \hat{x}(t_i + \Theta_i(\varepsilon)) + \varepsilon \sum_{j=1}^{r_i} c_j^i v_{ji} + o(\varepsilon). \end{aligned}$$

Let $\bar{\varepsilon} < \min \{\bar{\varepsilon}_i\}$ be such that $\sup_{\varepsilon \in [0, \bar{\varepsilon}]} \{t_i + \Theta_i(\varepsilon)\} < \inf_{\varepsilon \in [0, \bar{\varepsilon}]} \{t_{i+1} - \Phi_{i+1}(\varepsilon)\}$. Let $V = V_1 \times V_2 \times \dots \times V_p$; V is neighborhood of 0 in $(\mathbb{R}^+)^s$. We define $\mu: V \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$ by

$$\mu(c, \varepsilon)(t) = \begin{cases} \nu_i(c^i, \varepsilon)(t), & t \in [t_i - \Phi_i(\varepsilon), t_i + \Theta_i(\varepsilon)], \\ \hat{u}(t), & \text{otherwise.} \end{cases}$$

Let us prove by induction that

$$(1.9) \quad \begin{aligned} & S(t_i + \Theta_i(\varepsilon), t_0, \xi_0, \mu(c, \varepsilon)) = \hat{x}(t_i + \Theta_i(\varepsilon)) \\ &+ \varepsilon \sum_{j=1}^i \sum_{s=1}^{r_j} c_s^j (\hat{S}(t_i + \Theta_i(\varepsilon), t_j, \cdot))_* v_{sj} + o(\varepsilon). \end{aligned}$$

For $i = 1$, (1.9) is proved by (1.8). Let $a_i(\varepsilon) = t_i + \Theta_i(\varepsilon)$ and $b_i(\varepsilon) = t_i - \Phi_i(\varepsilon)$. Then

$$\begin{aligned} & S(a_{i+1}(\varepsilon), t_0, \xi_0, \mu(c, \varepsilon)) = S(a_{i+1}(\varepsilon), b_{i+1}(\varepsilon), \\ & \quad \hat{S}(b_{i+1}(\varepsilon), a_i(\varepsilon), S(a_i(\varepsilon), t_0, \xi_0, \mu(c, \varepsilon))), \mu(c, \varepsilon)) \\ &= S\left(a_{i+1}(\varepsilon), b_{i+1}(\varepsilon), \hat{S}\left(b_{i+1}(\varepsilon), a_i(\varepsilon), \hat{x}(a_i(\varepsilon))\right.\right. \\ & \quad \left.\left.+ \varepsilon \left[\sum_{j=1}^i \sum_{s=1}^{r_j} c_s^j (\hat{S}(a_i(\varepsilon), t_j, \cdot))_* v_{sj} \right] + o(\varepsilon) \right), \nu_{i+1}(c^{i+1}, \varepsilon)\right) \\ &= S\left(a_{i+1}(\varepsilon), b_{i+1}(\varepsilon), \hat{x}(b_{i+1}(\varepsilon))\right. \\ & \quad \left.+ \varepsilon \left[\sum_{j=1}^i \sum_{s=1}^{r_j} c_s^j (\hat{S}(b_{i+1}(\varepsilon), t_j, \cdot))_* v_{sj} \right] \right. \\ & \quad \left.+ o(\varepsilon), \nu_{i+1}(c^{i+1}, \varepsilon)\right) \\ &= S(a_{i+1}(\varepsilon), b_{i+1}(\varepsilon), \hat{x}(b_{i+1}(\varepsilon)), \nu_{i+1}(c^{i+1}, \varepsilon)) \\ & \quad + \varepsilon \left[\sum_{j=1}^i \sum_{s=1}^{r_j} c_s^j (\hat{S}(a_{i+1}(\varepsilon), t_j, \cdot))_* v_{sj} \right] + o(\varepsilon) \\ &= \hat{x}(t_{i+1} + \Theta_{i+1}(\varepsilon)) + \varepsilon \left[\sum_{j=1}^{i+1} \sum_{s=1}^{r_j} c_s^j (\hat{S}(a_{i+1}(\varepsilon), t_j, \cdot))_* v_{sj} \right] + o(\varepsilon). \end{aligned}$$

If $t_p = T$, then $\Theta_p(\varepsilon) = 0$ since there are no right variations at T ; hence (1.9) proves the theorem for $i = p$. Otherwise, if $t_p < T$, we set $t_{p+1} = T$, $\Theta_{p+1}(\varepsilon) \equiv \Phi_{p+1}(\varepsilon) \equiv 0$, and (1.9) proves the theorem for $i = p + 1$. \square

2. C^∞ time-independent control systems. In this section we study the properties of the variational cone in the case of C^∞ time-independent control systems. Note that

any control system Σ can be considered to be time independent by considering the time as another state variable. Hence the results of this section can be adapted to the time-dependent systems that are C^∞ with respect to both time and state. For the C^∞ manifolds we use the following notations: $\mathcal{V}(M)$ is the Lie algebra of C^∞ vector fields on M . If $f \in \mathcal{V}(M)$, $(t, \xi) \mapsto \exp tf \cdot \xi$ denotes the local flow of f and $ad_f: \mathcal{V}(M) \rightarrow \mathcal{V}(M)$, $g \mapsto ad_f g \equiv [f, g]$ denotes the Lie derivation in $\mathcal{V}(M)$ with respect to f . If F is a subset of $\mathcal{V}(M)$, $\text{Lie}(F)$ denotes the Lie subalgebra of $\mathcal{V}(M)$ generated by F and $(\text{Lie}(F))'$ denotes the derived algebra of $\text{Lie}(F)$ defined by

$$(\text{Lie}(F))' = \text{span} \{[X, Y]: X, Y \in \text{Lie}(F)\}.$$

We consider the control system Σ :

$$\dot{x} = f(x, u).$$

In addition to the hypotheses of § 1, we assume the following:

(A5) The manifold M is C^∞ .

(A6) $f: M \times \mathbb{R}^m \rightarrow TM$ is such that for each $\omega \in \mathbb{R}^m$, $f(\cdot, \omega)$ is a C^∞ vector field.

(A7) If $u: [a, b] \rightarrow \mathbb{R}^m$ belongs to \mathcal{U} , then $\forall \tau$ the control map $u_\tau: [a + \tau, b + \tau] \rightarrow \mathbb{R}^m$, defined by $u_\tau(t) = u(t - \tau)$, belongs to \mathcal{U} .

Assumption (A7) ensures that if one has an admissible control, then one can use it at any time, so that the trajectories of Σ can always be thought of as trajectories with initial time equal to 0. For this reason from now on we will not specify the initial time. For example, $S(t, \xi, u)$ will denote the value at time t of the solution of Σ relative to the control u starting at ξ . Notice that because Σ is time independent, u_τ fulfills assumption (A2) if and only if u does.

Let Ω be the set of all the values taken by the control maps, i.e., $\Omega = \{u(t) \in \mathbb{R}^m: u \in \mathcal{U}, t \in I_u\}$. We associate with Σ the family $F = \{f(\cdot, \omega): \omega \in \Omega\} \subseteq \mathcal{V}(M)$. The trajectories of Σ relative to the piecewise-constant controls are contained in the orbits of F . The local flows of the vector fields in F generate a pseudogroup of local omeomorphisms:

$$G = \{\exp t_1 g_1 \cdots \exp t_k g_k: t_i \in \mathbb{R}, g_i \in F\}.$$

It is known that $N(\xi_0) = \{\phi(\xi_0): \phi \in G\}$ is a C^∞ immersed, connected submanifold of M [19]. $N(\xi_0)$ is the set of points that can be reached from ξ_0 by means of the orbits of F . Let $N(\xi_0, t)$ be the subset of $N(\xi_0)$ defined by

$$N(\xi_0, t) = \left\{ \phi(\xi_0): \phi \in G, \sum_{i=1}^k t_i = t \right\}.$$

$N(\xi_0, t)$ is a possibly disconnected, integral manifold of the distribution

$$\Delta^0 = \{g_* \phi \circ g^{-1} - h_* \chi \circ h^{-1}: g, h \in G, \phi, \chi \in F\}.$$

Δ^0 always contains the distribution

$$J^0(F) = \text{span} \left\{ \sum a_i X_i + Z: X_i \in F, a_i \in \mathbb{R}, \sum a_i = 0, Z \in (\text{Lie}(F))' \right\},$$

and Δ^0 coincides with J^0 if either $J^0(F)$ has constant dimension or the family F is analytic; see [4], [6] and the references therein.

In [4] it is proved that, whatever \mathcal{U} is with the properties A_1, \dots, A_7 , the reachable set $R(\xi_0, t)$ is contained in $N(\xi_0, t)$.

DEFINITION 2.1. The relative interior, $\text{int}_{\text{rel}} R(\xi_0, t)$, of $R(\xi_0, t)$ is the set of interior points of $R(\xi_0, t)$ in the topology of $N(\xi_0, t)$.

Remark 2.2. If \mathcal{U} contains the set of piecewise-constant controls with values in Ω and $R(\xi_0, t) \neq \emptyset$, then $\text{int}_{\text{rel}} R(\xi_0, t)$ is not empty if either F is analytic or $J^0(F)$ has constant dimension [20].

DEFINITION 2.3. The control system is *weakly locally controllable* along a trajectory \hat{x} (or, if the system can be understood, \hat{x} is weakly locally controllable) if and only if for each $t \in (0, T]$

$$\hat{x}(t) \in \text{int}_{\text{rel}} R(\xi_0, t).$$

By Theorem 1.10, if for each t positive $\mathcal{K}(t) = \Delta^0(\hat{x}(t))$, then \hat{x} is weakly locally controllable. For example, this is the case if $\mathcal{K}(0) = \Delta^0(\xi_0)$.

A variation at t or, more generally, an element $w \in \mathcal{K}(t)$ can be thought of as the value at $\hat{x}(t)$ of a vector field Y that is not uniquely determined. We will say that a vector field $Y \in \mathcal{V}(M)$ defines a variation at time t (an element of $\mathcal{K}(t)$) if and only if $Y(\hat{x}(t))$ is a variation at t (is an element of the variational cone at time t). In this sense we can say that a subset of $\mathcal{V}(M)$ is either a set of variations at t or a subset of the variational cone at t .

The subspaces contained in the variational cone are particularly interesting. As a matter of fact, if \hat{x} is time optimal, then the adjoint covector given by the Pontryagin maximum principle must be orthogonal to them. The subspaces of $\mathcal{V}(M)$ contained in the variational cone for each t in an interval give rise to further tangent directions. Namely, as a consequence of the properties of the transport along the reference flow, we have

THEOREM 2.4. *Let us suppose that the reference control is constant in $(t_0, t_1]$, i.e., $\hat{u}(t) = \omega \forall t \in (t_0, t_1]$, and let $f_0 = f(\cdot, \omega)$. If \mathcal{D} is a subspace of $\mathcal{V}(M)$ such that*

$$(2.1) \quad Z(\hat{x}(t)) \in \mathcal{K}(t) \quad \forall Z \in \mathcal{D} \text{ and } \forall t \in (t_0, t_1],$$

then for all $Z \in \mathcal{D}$

$$ad_{f_0}^i Z(\hat{x}(t)) \in \mathcal{K}(t) \quad \forall i \in \mathbb{N} \text{ and } \forall t \in (t_0, t_1].$$

Proof. Let $t_0 < t \leq t_1$; by definition $(\exp \tau f_0)_* Z(\hat{x}(t - \tau)) \in \mathcal{K}(t)$, $\forall \tau \in [0, t - t_0]$. Hence the subspace $H = \text{co} \{(\exp \tau f_0)_* Z(\hat{x}(t - \tau)) : Z \in \mathcal{D}, \tau \in [0, t - t_0]\}$ is contained in $\mathcal{K}(t)$. The C^∞ curve $\tau \mapsto (\exp \tau f_0)_* Z(\hat{x}(t - \tau))$ belongs to H , so that its derivatives at $\tau = 0$ belong to H and hence to $\mathcal{K}(t)$. However, the i th derivative at $\tau = 0$ is $ad_{f_0}^i Z(\hat{x}(t))$, so that the statement is proved. \square

Hypothesis (A7) implies that if the reference trajectory is stationary, then the set of variations at t does not depend on t , so that the variational cone is increasing with t . Hence in this case the maximal subspace contained in the variational cone at any positive time must be invariant under the Lie derivative with respect to f_0 . The variations, and hence the variational cone, of a stationary trajectory have further specific properties that are consequences of the fact that the control variation of Definition 1.2 can be chosen independently on γ . The properties we are going to prove are properties of variations of a stationary trajectory of a time-independent control system that satisfy hypothesis (A7), and they also hold if the control system is only C^1 . For variations of a stationary trajectory there is no distinction between right and left variations, and any variations can be thought of as a variation at any time. For this reason we omit the time at which a variation occurs and we consider each variation as a right variation.

PROPOSITION 2.5. *If v is a variation of order α relative to a stationary trajectory, it is a variation of any order greater than α .*

Proof. By Definition 1.2 there exists a continuous control variation $\eta(c, \varepsilon)$ such that in a chart at ξ_0

$$S(\varepsilon, \xi_0, \eta(c, \varepsilon)) = \xi_0 + \varepsilon^\alpha cv + o(\varepsilon^\alpha).$$

Let $\beta > 1$; ε^β is smaller than ε for $0 < \varepsilon < 1$. If $\bar{\eta}(c, \varepsilon)$ coincides with the control $\eta(c, \varepsilon^\beta)$ in the interval $[0, \varepsilon^\beta]$ and with the reference control in the interval $[\varepsilon^\beta, \varepsilon]$, then

$$S(\varepsilon, \xi_0, \bar{\eta}(c, \varepsilon)) = \xi_0 + \varepsilon^{\alpha\beta} cv + o(\varepsilon^{\alpha\beta}),$$

so that the proposition is proved. \square

Notice that for a nonstationary trajectory the variation's order cannot be enlarged in an analogous way. In fact, it may not be possible to enlarge the length $\varepsilon\gamma$ of the interval of possible starting times of control variations.

Lemma 1.15 implies the following:

COROLLARY 2.6. *The set of variations relative to a stationary trajectory is convex.*

3. Affine control systems. In this section we study the properties of the variational cone when the control system is a C^∞ system affine with respect to the control and the reference control is constant. Without loss of generality we can suppose that $\hat{u}(t) \equiv 0$. The results obtained can also be used when the reference control is a piecewise-constant one. In fact, if we restrict the reference couple to a subinterval, we obtain a variational cone that is contained in the original one.

Let Ω be an assigned subset of \mathbb{R}^m such that $0 \in \Omega$ and $\text{span } \Omega = \mathbb{R}^m$. To each family $\mathbf{f} = (f_0, f_1, \dots, f_m)$ of C^∞ vector fields on a manifold M we associate the affine control process $(\Sigma_{\mathbf{f}}, \Omega)$ on M , where $\Sigma_{\mathbf{f}}$ is defined by

$$\dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$$

and the control u belongs to the class of the piecewise-constant maps with values in Ω . We will deal with controls defined on subintervals of a given interval $[0, T]$, so that we will take as admissible controls the set \mathcal{U} of piecewise-constant maps from the subintervals of $[0, T]$ to Ω . The set \mathcal{U} can be immersed in $L^1([0, T], \mathbb{R}^m)$ by extending each $u \in \mathcal{U}$ with the zero value. We choose Π equal to the topology induced on \mathcal{U} by the L^1 topology.

$(\Sigma_{\mathbf{f}}, \Omega)$ belongs to the class of systems considered in the previous sections. We will attach the subscript \mathbf{f} to any object defined in §§ 1 and 2 for a general system when it is referred to the system $\Sigma_{\mathbf{f}}$. For example, $S_{\mathbf{f}}(t, \xi, u)$ will denote the value at time t of the solution of $\Sigma_{\mathbf{f}}$ relative to the control u , starting at ξ . For this class of systems the distribution $J_{\mathbf{f}}^0$ defined in § 2 coincides with the distribution

$$\mathcal{S}_{\mathbf{f}} = \text{Lie} \{ad_{f_0}^i f_i, i = 1, \dots, m, j = 0, 1, \dots, \}.$$

Let $\hat{u} \equiv 0$ be the reference control, and let $\hat{x}_{\mathbf{f}}(t) = S_{\mathbf{f}}(t, \xi_0, 0)$, $t \in [0, T]$ be the reference trajectory; $\mathcal{K}_{\mathbf{f}}(t)$ is the variational cone of the reference couple $(\hat{x}_{\mathbf{f}}, 0)$. A vector v is a right variation of order α at t if there exists a continuous map $(\gamma, c, \varepsilon) \mapsto \eta(\gamma, c, \varepsilon)$ such that

$$\exp(-\varepsilon\gamma - \varepsilon)f_0 \cdot S_{\mathbf{f}}(\varepsilon, \hat{x}_{\mathbf{f}}(t + \varepsilon\gamma), \eta(\gamma, c, \varepsilon)) = \hat{x}_{\mathbf{f}}(t) + \varepsilon^\alpha cv + o(\varepsilon^\alpha).$$

Note that because η is continuous, its L^1 norm $\|\eta(\gamma, c, \varepsilon)\|_{L^1}$ goes to 0 with ε . In fact, we can suppose that the support of the map $\eta(\gamma, c, \varepsilon)$ is contained in the interval

$[0, \varepsilon]$; this implies that

$$\begin{aligned} \|\eta(\gamma, c, \varepsilon)\|_{L^1} &= \int_0^\varepsilon \|\eta(\gamma, c, \varepsilon)(s)\| ds \leq \int_0^\varepsilon \|\eta(\gamma, c, \varepsilon)(s) - \eta(\gamma, c, 0)(s)\| ds \\ &\quad + \int_0^\varepsilon \|\eta(\gamma, c, 0)(s)\| ds \\ &\leq \|\eta(\gamma, c, \varepsilon) - \eta(\gamma, c, 0)\|_{L^1} + \int_0^\varepsilon \|\eta(\gamma, c, 0)(s)\| ds. \end{aligned}$$

The flow of the control system S_t , as a map from a subset of $[0, T] \times M \times L_1([0, T], \mathbb{R}^m)$ to M , is continuous. Therefore, the variations at τ can be studied in any chart at $\hat{x}_t(\tau)$. In other words, the trajectory's variations are contained in a prescribed chart for ε small.

The following approximation result shows that the variations at τ depend in general on only the germs of the f_i 's at $\hat{x}_t(\tau)$. Without loss of generality set $\tau = 0$. Let us choose a coordinate system at $\xi_0 \in M$, and let us use the same notation for the system Σ_t and for its coordinate representation. The following property holds:

PROPOSITION 3.1. *Let Σ_Φ denote the affine control system obtained by substituting in the system Σ_t at each vector field f_i its Taylor approximation at ξ_0 of order k , and let $\alpha < k + 1$. If $\eta : [0, \bar{\gamma}] \times [0, \bar{c}] \times [0, \bar{\varepsilon}] \rightarrow \mathcal{U}$ is a continuous map such that*

$$(3.1) \quad (\|\eta(\gamma, c, \varepsilon)\|_{L^1})^{k+1} = o(\varepsilon^\alpha),$$

then

$$S_t(\varepsilon, \hat{x}_t(\varepsilon\gamma), \eta(\gamma, c, \varepsilon)) = \hat{x}_t(\varepsilon\gamma + \varepsilon) + \varepsilon^\alpha cv + o(\varepsilon^\alpha)$$

if and only if

$$S_\Phi(\varepsilon, \hat{x}_\Phi(\varepsilon\gamma), \eta(\gamma, c, \varepsilon)) = \hat{x}_\Phi(\varepsilon\gamma + \varepsilon) + \varepsilon^\alpha cv + o(\varepsilon^\alpha).$$

Proof. Let I be a compact interval containing 0 such that $\hat{x}_\Phi(t)$ is defined $\forall t \in I$. The Gronwall's lemma implies that there exist two positive constants H, C such that if $\|u\|_{L^1} < H$, then

$$(3.2) \quad \|S_\Phi(t, \xi_0, u)\| \leq \|\hat{x}_\Phi(t)\| + C\|u\|_{L^1} \quad \forall t \in I.$$

Let $G(x)$ denote the $(m+1) \times n$ matrix whose columns are the vector fields $f_i(x)$, $i = 0, 1, \dots, m$, and let $G_k(x)$ be its Taylor approximation of order k . Let $u \in \mathcal{U}$, $\|u\|_{L^1} < H$; we set $v(t)$ equal to the transpose of $(1, u_1(t), \dots, u_m(t))$.

$$\begin{aligned} \|S_t(t, \xi_0, u) - S_\Phi(t, \xi_0, u)\| &= \left\| \int_0^t (G(S_t(\tau, \xi_0, u)) - G_k(S_\Phi(\tau, \xi_0, u))) \cdot v(\tau) d\tau \right\| \\ &= \left\| \int_0^t (G(S_t(\tau, \xi_0, u)) - G(S_\Phi(\tau, \xi_0, u)) + G(S_\Phi(\tau, \xi_0, u)) \right. \\ &\quad \left. - G_k(S_\Phi(\tau, \xi_0, u))) \cdot v(\tau) d\tau \right\| \\ &\leq \int_0^t L\|v(\tau)\| \|S_t(\tau, \xi_0, u) - S_\Phi(\tau, \xi_0, u)\| d\tau \\ &\quad + \int_0^t N\|v(\tau)\| \|(S_\Phi(\tau, \xi_0, u))^{(k+1)}\| d\tau \\ &\leq \int_0^t L\|v(\tau)\| \|S_t(\tau, \xi_0, u) - S_\Phi(\tau, \xi_0, u)\| d\tau \\ &\quad + N(t+H) \left(\sup_{\tau \in [0, t]} \|S_\Phi(\tau, \xi_0, u)\|^{k+1} \right) \end{aligned}$$

for some positive constants N, L . Gronwall's inequality implies

$$\|S_t(t, \xi_0, u) - S_\Phi(t, \xi_0, u)\| \leq B \left(\sup_{\tau \in [0, t]} \|S_\Phi(\tau, \xi_0, u)\|^{k+1} \right).$$

By (3.1) if ε is sufficiently small, $\|\eta\|_{L^1} < H$. Let $\bar{\eta}(\gamma, c, \varepsilon)$ be the control map defined by

$$\bar{\eta}(\gamma, c, \varepsilon)(t) = \begin{cases} 0, & t \in [0, \varepsilon\gamma], \\ \eta(\gamma, c, \varepsilon)(t), & t \in (\varepsilon\gamma, \varepsilon + \varepsilon\gamma], \end{cases}$$

$\|\eta\|_{L^1} = \|\bar{\eta}\|_{L^1}$, and

$$\begin{aligned} & \|S_\Phi(\varepsilon, \hat{x}_\Phi(\varepsilon\gamma), \eta(\gamma, c, \varepsilon)) - \hat{x}_\Phi(\varepsilon\gamma + \varepsilon) - S_t(\varepsilon, \hat{x}_t(\varepsilon\gamma), \eta(\gamma, c, \varepsilon)) + \hat{x}_t(\varepsilon\gamma + \varepsilon)\| \\ & \leq \|\hat{x}_\Phi(\varepsilon\gamma + \varepsilon) - \hat{x}_t(\varepsilon\gamma + \varepsilon)\| + \|S_t(\varepsilon\gamma + \varepsilon, \xi_0, \bar{\eta}(\gamma, c, \varepsilon)) \\ & \quad - S_\Phi(\varepsilon\gamma + \varepsilon, \xi_0, \bar{\eta}(\gamma, c, \varepsilon))\| \\ & \leq B \left(\sup_{\tau \in [0, 2\varepsilon]} \|\hat{x}_\Phi(\tau)\|^{k+1} + \sup_{\tau \in [0, 2\varepsilon]} \|S_\Phi(\tau, \xi_0, \bar{\eta})\|^{k+1} \right). \end{aligned}$$

$\|\hat{x}_\Phi(\tau)\|$ goes to zero at least as τ , so that $\sup_{\tau \in [0, 2\varepsilon]} \|\hat{x}_\Phi(\tau)\|^{k+1} = o(\varepsilon^\alpha)$; moreover, by (3.1) and (3.2) we get $\sup_{\tau \in [0, 2\varepsilon]} \|S_\Phi(\tau, 0, \bar{\eta})\|^{k+1} = o(\varepsilon^\alpha)$, and the statement is proved. \square

Remark 3.2. Let v be a variation of order α such that the L^1 -norm of its control variation goes to 0 as ε^r . v remains a variation of order α if the vector fields f_i 's are perturbed with terms of order greater than $\max\{\alpha, \alpha/r\}$.

Notice that if Ω is bounded, then $\|\eta(\gamma, c, \varepsilon)\|_{L^1} \leq L\varepsilon$ for some constant L , so that the following corollary holds.

COROLLARY 3.3. *If Ω is bounded, v is a variation of order α for $(\hat{x}_t, 0)$ at time 0 if and only if it is a variation of order α at time 0 for $(\hat{x}_\Phi, 0)$ for each $k > \alpha - 1$.*

A way to compare variations that occur at different times is to bring each variation back to time 0 by using the reference flow, that is, $v \in T_{\hat{x}_t(t)}M$ is a variation at time t if and only if

$$\exp(-t - \varepsilon\gamma - \varepsilon)f_0 \cdot S_t(\varepsilon, \hat{x}_t(t + \varepsilon\gamma), \eta(\gamma, c, \varepsilon)) = \xi_0 + \varepsilon^\alpha c \exp(-tf_0)_* v + o(\varepsilon^\alpha).$$

This point of view is better understood by looking at the pullback system introduced in [3]. This system describes the behavior of the trajectories of Σ_t that lie near the reference trajectory relative to the reference trajectory itself. Hence it can be considered to be the “right” system for studying the variations of the reference couple. Let us recall from [3] the definition of a pullback system.

By the properties of differential equations it follows that there is a neighborhood U of ξ_0 and a neighborhood I of 0 in \mathbb{R} containing $[0, T]$ such that $\exp tf_0 \cdot \xi$ is defined for each t in I and each ξ in U . Therefore, we can define for any $f \in \mathcal{V}(M)$ the time-dependent vector field

$$f^*: I \times U \rightarrow TU \subset TM, \quad f^*(t, \xi) \equiv f^*(t)(\xi) = \exp(-tf_0)_* \cdot f(\exp tf_0 \cdot \xi).$$

If u belongs to a sufficiently small neighborhood of 0 in $L^1([0, T], \Omega)$, the map

$$t \mapsto y(t, \xi_0, u) \equiv \exp(-tf_0) \cdot S_t(t, \xi_0, u)$$

is defined and it satisfies the time-dependent control system on U :

$$\dot{y}(t) = \sum_{i=1}^m u_i(t) f_i^*(t, y(t)), \quad y(0) = \xi_0.$$

The time-dependent vector field f_i^* may be viewed as a C^∞ vector field on $M^* = I \times U$. If we set $\mathbf{f}^* = \{\partial/\partial t, f_1^*, \dots, f_m^*\}$, the pullback system of the system Σ_r is the system Σ_{r^*} on M^* given by

$$\dot{y}^* = \frac{\partial}{\partial t} + \sum_{i=1}^m u_i f_i^*(y^*).$$

Notice that there is an isomorphism between the solutions of Σ_{r^*} and the ones of the restriction of Σ_r to the neighborhood $V = \{\exp tf_0 \cdot \xi : \xi \in U, t \in I\}$ of the reference trajectory. The solutions of the two systems are linked by the relation

$$(3.3) \quad S_{r^*}(t, (\tau, \xi_0), u) = (t + \tau, \exp(-(t + \tau)f_0) \cdot S_r(t, \exp \tau f_0 \cdot \xi_0, u)).$$

In what follows we shall identify U with $\{0\} \times U$, so that $\xi_0 \equiv (0, \xi_0)$, $N_{r^*}(\xi_0, 0)$ is contained in $N_r(\xi_0, 0)$, $N_{r^*}(\xi_0, t) = \{t\} \times N_{r^*}(\xi_0, 0)$ and $\Delta_{r^*}^0(\hat{x}_{r^*}(t)) = \Delta_{r^*}^0(\xi_0)$. Notice that Σ_{r^*} evolves on $I \times N_{r^*}(\xi_0, 0)$ and that the reference flow is a translation. In some sense Σ_{r^*} transforms an open neighborhood of the reference trajectory in the product $I \times U$.

Since Σ_{r^*} reproduces only the trajectories of Σ_r lying near the reference trajectory, $N_{r^*}(\xi_0, 0)$ may be not open in $N_r(\xi_0, 0)$, but it is if $\Delta_{r^*}^0(\xi_0) = \Delta_r^0(\xi_0)$. In particular, this is the case if $\mathcal{S}_r(\xi_0) = \Delta_r^0(\xi_0)$; in fact, $\mathcal{S}_{r^*} = \{h^* : h \in \mathcal{S}_r\}$. In any case the variations are generated by small (with respect to the L^1 norm) controls, so that from (3.3) it follows that a vector field h is such that $h(\hat{x}_r(t))$ is a variation at t for $(\hat{x}_r, 0)$ if and only if $h^*(t, \xi_0)$ is a variation at t for $(\hat{x}_{r^*}, 0)$. In particular, $v \in \mathcal{H}_r(0)$ if and only if $v \in \mathcal{H}_{r^*}(0)$. In terms of pullback the vector $v \in \Delta_r^0(\xi_0)$ is a right variation of order α at t if there exists a continuous map $(\gamma, c, \varepsilon) \mapsto \eta(\gamma, c, \varepsilon)$ such that

$$S_{r^*}(\varepsilon, (t + \varepsilon\gamma, \xi_0), \eta(\gamma, c, \varepsilon)) = (t + \varepsilon\gamma + \varepsilon, \xi_0 + \varepsilon^\alpha cv + o(\varepsilon^\alpha)).$$

Each f_i , $i = 1, \dots, m$, provides a variation of order 1 at each time (see Example 1.3). Moreover, the Campbell–Hausdorff formula gives any trajectory of Σ_{r^*} relative to a piecewise-constant control as the trajectory of a vector field whose asymptotic expansion (modulo $\partial/\partial t$) is an element of \mathcal{S}_{r^*} . Noting that $\mathcal{S}_{r^*}(\xi_0) = \mathcal{S}_r(\xi_0)$, we shall keep the elements of $\mathcal{S}_r(\xi_0)$ as candidates to be variations at 0.

The main result in [3] provides a sufficient condition for (Σ_r, Ω) to be weakly locally controllable along \hat{x}_r when Ω is a neighborhood of the origin in \mathbb{R}^m . Namely, some elements of \mathcal{S}_r are pointed out as possible obstructions, and some conditions are given to neutralize them. In this paper we use the same ideas to construct variations at 0. The result is that different types of neutralization can be used for the same system; see Example 5.2. To make the above ideas more precise and to state the main result we need the notations introduced in [3], [18].

Let $\text{Lie } \mathbf{X}$ be the free Lie algebra on \mathbb{R} generated by the noncommutative indeterminates $\mathbf{X} = \{X_0, \dots, X_m\}$. \mathcal{S} will denote the ideal of $\text{Lie } \mathbf{X}$ generated by X_1, \dots, X_m . Substituting f_i for X_i in any element $\chi \in \text{Lie } \mathbf{X}$, we obtain a vector field that will be denoted by χ_r . For any subset A of $\text{Lie } \mathbf{X}$, A_r will denote the subset of $\text{Lie } \mathbf{f}$ given by $A_r = \{\chi_r : \chi \in A\}$. By means of a set $\mathbf{l} = (l_0, \dots, l_m)$ of integers we define a weight on $\text{Lie } \mathbf{X}$ that will induce a weight on $\text{Lie } \mathbf{f}$. Let Λ be a bracket in \mathcal{S} . We denote the length of Λ with respect to X_i (i.e., the number of times that X_i appears in Λ) by $|\Lambda|_i$. The weight of Λ is defined by

$$\|\Lambda\|_{\mathbf{l}} = \sum_{i=0}^m l_i |\Lambda|_i, \quad \|0\|_{\mathbf{l}} = 0,$$

and the subspace of \mathcal{S}_r of the elements of weight not greater than i is given by

$$V_i^r = \text{span} \{\Lambda_r : \Lambda \in \mathcal{S}, \|\Lambda\|_{\mathbf{l}} \leq i\}.$$

An element $\chi \in \mathcal{S}$ is called *l-homogeneous* if it is a linear combination of brackets with the same weight, which will be called the *weight* of χ . Following H. J. Sussmann, we say that an l-homogeneous element $\chi \in \mathcal{S}$ is *l-neutralized* for Σ_r at ξ_0 if χ_r is a linear combination at ξ_0 of brackets with less weight. In other words, χ is l-neutralized if there is an $i < \|\chi\|_1$ such that $\chi_r(\xi_0) \in V_i^1(\xi_0)$. For more details and examples see [3], [18].

To identify the elements of $\text{Lie } \mathfrak{f}$ that give rise to variations, we introduce the set of obstructions relative to a weight **l**; see [3]. Let

$$\mathcal{B} = \text{span} \{ \Lambda \in \mathcal{S} : |\Lambda|_0 \text{ is odd, } |\Lambda|_i \text{ is even, } i = 1, \dots, m \},$$

$$\mathcal{B}_S^1 = \{ \chi \in \mathcal{B} : \chi \text{ is symmetric w.r.t. those } X_i \text{'s that have the same weight, } i \neq 0 \}.$$

In other words, the elements of \mathcal{B}_S^1 are the fixed elements of the automorphisms of \mathcal{B} generated by $\mu_{ij} : X_i \mapsto X_j, \forall i, j$ such that $l_i = l_j$. For example, if $m = 3, l_0 = 0, l_1 = l_2 = 1, l_3 = 3$, then $ad_{X_1}^2 X_0 + ad_{X_2}^2 X_0$ and $ad_{X_3}^2 X_0$ belong to \mathcal{B}_S^1 , but $ad_{X_1}^2 X_0$ does not.

The set of obstructions relative to the weight **l** is the set

$$\mathcal{B}_l^* = \text{Lie}(X_0, \mathcal{B}_S^1) \cap \mathcal{S}.$$

DEFINITION 3.4. A set **l** = (l_0, \dots, l_m) of integers will be called a *set of admissible weights* for Ω if and only if for each $\omega = (\omega_1, \dots, \omega_m) \in \Omega$ and each $\varepsilon \in (0, 1)$

$$(3.4) \quad (\varepsilon^{l_1 - l_0} \omega_1, \dots, \varepsilon^{l_m - l_0} \omega_m) \in \Omega.$$

Notice that if Ω is a polydisk $H_p = \{(\omega_1, \dots, \omega_m) \in \mathbb{R}^m : |\omega_i| \leq \rho_i, i = 1, \dots, m\}$, possibly some $\rho_i = +\infty$, then (3.4) is equivalent to requiring that if ρ_i is finite, then $l_i \geq l_0$.

The following result states that an element of \mathcal{S}_r defines a variation at time zero if every obstruction with lower or equal weight is neutralized at ξ_0 by means of a set of admissible weights.

THEOREM 3.5. Let Ω be the polydisk $H_p = \{(\omega_1, \dots, \omega_m) \in \mathbb{R}^m : |\omega_i| \leq \rho_i, i = 1, \dots, m\}$, possibly some $\rho_i = +\infty$, and let Φ be a bracket in \mathcal{S} . If there exists a set of admissible nonnegative weights, **l** = (l_0, \dots, l_m) , such that

$$(3.5) \quad \text{each } \Lambda \in \mathcal{B}_l^* \text{ s.t. } \|\Lambda\|_1 \leq \|\Phi\|_1 \text{ is l-neutralized at } \xi_0$$

(i.e., Λ_r is a linear combination at ξ_0 of elements in \mathcal{S}_r with lower weight), then $\Phi_r(\xi_0)$ is a regular variation for the pair $(\hat{x}_r, 0)$. Moreover, if all the l_i 's are positive, then the order of the variation is $\|\Phi\|_1 / l_0$.

The proof of the theorem is based on the properties of suitable graded approximating systems introduced in [3]. In § 4 we shall investigate the links between the variations of the system and those of the approximating system, and we will provide the proof of the theorem. Notice that the approximation result in § 4 is different from the one stated in Proposition 3.1. In fact, the result in Proposition 3.1 links the order of a single variation to the order of the Taylor approximation of the fields. In § 4 we state that suitable subspaces of the set of variations of a first-order graded approximating system are also variations of the original system.

Theorem 3.5 provides subspaces contained in $\mathcal{K}_r(0)$; in fact, if a bracket $\Phi \in \mathcal{S}$ satisfies the hypotheses of the theorem, the same is true for each bracket with the same **l**-weight. Hence it can be used to derive by means of Corollary 1.11 and Remark 1.14 either sufficient conditions of weak local controllability or necessary conditions for the couple $(\hat{x}_r, 0)$ to be optimal. The applications to the local controllability property will be described in details in § 5. Here we limit ourselves to constructing subspaces of variations.

PROPOSITION 3.6. If Ω is a polydisk, the distributions

$$H_1 = \text{span} \{ f_j : j = 1, \dots, m \},$$

$$H_2 = \text{span} \{ [f_i, f_j] : i, j = 1, \dots, m \}$$

define two subspaces of regular variations at any time. If, moreover, $\Omega = \mathbb{R}^m$, then $\text{Lie}\{f_1, \dots, f_m\}$ defines a subspace of regular variations at any time.

Proof. Let $\mathbf{l} = (1, 1, \dots, 1)$: $\Lambda \in \mathcal{B}_1^*$ implies $\|\Lambda\|_1 \geq 3$. Hence if $\Phi \in \mathcal{S}$ is a bracket of length no greater than two, then (3.4) is fulfilled. This implies that for $i = 1, \dots, m$, $\pm f_i(\hat{x}(t))$ are regular variations of order 1 and $\pm[f_i, f_j](\hat{x}(t))$ are regular variations of order 2. If $\Omega = \mathbb{R}^m$, we can choose $\mathbf{l} = (1, 0, \dots, 0)$. $\Phi \in \text{Lie}\{f_1, \dots, f_m\}$ implies $\|\Phi\|_1 = 0$, meanwhile, $\Lambda \in \mathcal{B}_1^*$ implies $\|\Lambda\|_1 \geq 1$. Therefore, Φ satisfies the hypotheses of Theorem 3.5, and the proof is complete. \square

The above proposition and Theorem 2.4 lead to the following two corollaries.

COROLLARY 3.7. *For each positive t , the distribution*

$$\mathcal{C} = \text{span}\{ad_{f_0}^h f_i, ad_{f_0}^k [f_i, f_j]: h, k \geq 0, i, j \in (1, \dots, m)\}$$

defines a subspace contained in $\mathcal{H}_t(t)$.

COROLLARY 3.8. *If $\Omega = \mathbb{R}^m$, for each positive t the distribution*

$$\mathcal{Y} = \text{span}(ad_{f_0}^h Y: h \geq 0, Y \in \text{Lie}\{f_1, \dots, f_m\})$$

defines a subspace contained in $\mathcal{H}_t(t)$.

It is known that if $\hat{x}: [0, T] \rightarrow M$ is a time-optimal trajectory, then $\hat{x}(t) \in \partial R(\hat{x}(0), t)$ for each $t \in [0, T]$. Corollaries 3.7 and 3.8 and Remark 1.14 imply the following result:

PROPOSITION 3.9. *If the trajectory $t \mapsto \hat{x}_t(t) = \exp t f_0 \cdot \xi_0$ is time optimal in $[0, T]$, then $\mathcal{C}(\hat{x}_t(t)) \neq T_{\hat{x}_t(t)} M$ for each $t \in [0, T]$. Moreover, the adjoint covector $\lambda(t)$ of Remark 1.14 is orthogonal to $\mathcal{C}(\hat{x}_t(t))$. If $\Omega = \mathbb{R}^m$, then $\mathcal{Y}(\hat{x}_t(t)) \neq T_{\hat{x}_t(t)} M$ and the adjoint covector $\lambda(t)$ is orthogonal to $\mathcal{Y}(\hat{x}_t(t))$.*

Up to now we have given examples of subsets of $\text{Lie}\{f_0, \dots, f_m\}$ that define variations along the whole trajectory. Note that if $v = \Phi(x_t(t))$ is one of these variations, then in a chart at $x_t(t)$ there exist vectors near v that are variations at time near t . We will now provide an example of a vector field that defines a variation v at $t = 0$, but the set of variations at positive times is separated from v . This fact underlines once again the local nature of this kind of variation.

Example 3.10. Let us consider the following system on \mathbb{R}^3 :

$$\dot{x} = 1 + u, \quad \dot{y} = ux, \quad \dot{z} = u(x^3 y + y^2),$$

where $\xi_0 = (0, 0, 0)$ and $|u| \leq \frac{1}{2}$. The reference trajectory is given by $\hat{x}_t(t) = (t, 0, 0)$. Let $\chi = [ad_{f_0}^2 f_1, ad_{f_1}^2 f_0] = (2 - 6x)(\partial/\partial z) \pm \chi$ define variations at 0. In fact, they satisfy the hypotheses of Theorem 3.5 with $\mathbf{l} = (2, 3)$ (for details see [3]). To see that both $\pm\chi$ cannot define variations at positive times, consider $\bar{\xi} = \hat{x}_t(\bar{t})$, $\bar{t} \in (0, \frac{1}{3})$. If $\pm\chi$ were variations at \bar{t} , then they would be variations at 0 for the trajectory \hat{y}_t defined by $t \mapsto \exp t f_0 \cdot \bar{\xi}$. Since $\mathcal{C}(\bar{\xi}) = \text{span}\{\partial/\partial x, \partial/\partial y\}$ and $\chi(\bar{\xi}) = (2 - 6\bar{t})(\partial/\partial z)$, by Corollary 1.11 \hat{y}_t would be locally controllable. However, $ad_{f_1}^2 f_0 = -(2x^3 + 6xy - 2y)(\partial/\partial z)$, so that $ad_{f_1}^2 f_0(\hat{x}_t(\bar{t})) = -2\bar{t}^3(\partial/\partial z)$ does not belong to $\mathcal{C}(\hat{x}_t(\bar{t}))$. The results in [15] imply that for each $t > 0$ sufficiently small, $\hat{y}_t(t) = \exp t f_0 \cdot \bar{\xi}$ belongs to the boundary of $R(\bar{\xi}, t)$, a contradiction.

Slightly modifying the above example, we can show that if a vector field does not provide variations on an arc of trajectory, then its adjoint with respect to the reference field may not define an element of $\mathcal{H}_t(t)$. This shows that hypothesis (2.1) in Theorem 2.4 cannot be dropped.

Example 3.11. Let us consider the following system on \mathbb{R}^4 :

$$\dot{x} = 1 + u, \quad \dot{y} = ux, \quad \dot{z} = u(x^3 y + y^2), \quad \dot{w} = z + ux^3 y,$$

where $\xi_0 = (0, 0, 0, 0)$ and $|u| \leq \frac{1}{3}$. The reference trajectory is given by $\hat{x}_t(t) = (t, 0, 0, 0)$. Let $\chi = [ad_{f_0}^2 f_1, ad_{f_1}^2 f_0] = (2 - 6x)(\partial/\partial z) + (3x^2 - 4x)(\partial/\partial w)$. χ satisfies the hypothesis of

Theorem 3.5 at 0 with $\mathbf{l} = (2, 3)$, so that $\pm\chi$ define variations at 0. Hence for each positive t , $\mathcal{H}_t(t)$ contains $\pm(\exp tf_0)_*\chi(\xi_0) = \pm(2\partial/\partial z + 2t\partial/\partial w)$ and $\mathcal{C}(\hat{x}_t(t)) = \text{span}\{(\partial)/(\partial x), (\partial)/(\partial y)\}$. If $\pm ad_{f_0}\chi = \pm(6\partial/\partial z + (6-12x)(\partial/\partial w))$ defined an element of $\mathcal{H}_t(t)$, $\mathcal{H}_t(t)$ would be \mathbb{R}^4 for each positive t , so that the reference trajectory would be locally controllable. We shall show that points of type $(t, 0, 0, a^2)$ cannot be reached if t is sufficiently small.

Let us suppose that the control u is such that

$$x(t) = x(S_t(t, \xi_0, u)) = t, \quad y(t) = y(S_t(t, \xi_0, u)) = 0, \quad z(t) = z(S_t(t, \xi_0, u)) = 0.$$

We obtain

$$\begin{aligned} \int_0^\tau u(s)x^3(s)y(s) \, ds &= \frac{x^2(\tau)y^2(\tau)}{2} - \int_0^\tau y^2(s)x(s)(1+u(s)) \, ds \\ &= \frac{x^2(\tau)y^2(\tau)}{2} - \frac{y^3(\tau)}{3} - \int_0^\tau y^2(s)x(s) \, ds, \end{aligned}$$

so that

$$z(\tau) = \frac{x^2(\tau)y^2(\tau)}{2} - \frac{y^3(\tau)}{3} - \int_0^\tau y^2(s)x(s) \, ds + \int_0^\tau y^2(s)u(s) \, ds.$$

$y(t) = z(t) = 0$ implies

$$(3.6) \quad \int_0^t y^2(s)(x(s) - u(s)) \, ds = 0.$$

As a consequence, we obtain

$$\begin{aligned} w(t) = w(S_t(t, \xi_0, u)) &= \int_0^t z(s) \, ds - \int_0^t x(s)y^2(s) \, ds \\ &= \int_0^t \left(\frac{y^2(s)x^2(s)}{2} - \frac{y^3(s)}{3} \right) ds + \int_0^t \int_0^s y^2(\tau)(u(\tau) - x(\tau)) \, d\tau \, ds - \int_0^t x(s)y^2(s) \, ds \\ &= \int_0^t \left(\frac{y^2(s)x^2(s)}{2} - \frac{y^3(s)}{3} \right) ds + \int_0^t (t-s)y^2(s)(u(s) - x(s)) \, ds - \int_0^t x(s)y^2(s) \, ds \\ &= \int_0^t \left(\frac{y^2(s)x^2(s)}{2} - \frac{y^3(s)}{3} - sy^2(s)(u(s) - x(s)) - x(s)y^2(s) \right) ds, \end{aligned}$$

where the last equality is by (3.6). Since $\frac{2}{3}s \leq x(s) \leq \frac{4}{3}s$ and $-\frac{2}{9}s^2 \leq y(s) \leq \frac{2}{9}s^2$,

$$w(t) \leq \int_0^t \frac{y^2(s)}{2} \left(s^2 + s^2 + \frac{s}{3} + \frac{4}{3}s^2 - \frac{2}{3}s \right) ds \leq 0 \quad \text{for sufficiently small } t.$$

4. Graded approximating systems and variations. This section is essentially devoted to constructing subspaces of variations for the reference couple $(\hat{x}_t, 0)$ from the properties of graded approximating systems defined in [3]. As a consequence, we will derive the proof of Theorem 3.5. For the exact definition of a graded structure and its properties with respect to control systems, see [3] and the references therein. Here we give only a short summary of this topic.

Let $\mathbf{x} = (x^1, \dots, x^n)$ be a chart on a neighborhood U of $\xi_0 \in M$ such that $\mathbf{x}(\xi_0) = 0$ and $\mathbf{x}(U)$ is a ball centered at 0, and let $\mathbf{w} = (w^1, \dots, w^n)$ be a set of positive integers. The couple (\mathbf{x}, \mathbf{w}) defines a graded structure on U that will be called a *local graded structure* at ξ_0 and will be denoted by $(\mathbf{x}, U, \mathbf{w})$.

The graded structure is given by the dilations, i.e., maps δ_ε on U with values in U defined by

$$x^i \circ \delta_\varepsilon = \varepsilon^{w^i} x^i, \quad i = 1, \dots, n, \quad |\varepsilon| < 1.$$

In other words, we give the weight w^i to x^i . As a consequence, for each multi-index $\alpha = (\alpha_1, \dots, \alpha_n)$ the weight of the monomial $x^\alpha \equiv (x^1)^{\alpha_1} \dots (x^n)^{\alpha_n}$ is defined by $\mathcal{W}(x^\alpha) = \sum_{i=1}^n w^i \alpha_i$.

The *weight of a polynomial* (i.e., an element of the subalgebra \mathcal{D} of the algebra of C^∞ function on U generated by the coordinate functions x^1, \dots, x^n) is the greatest weight of the monomials contained in it. A polynomial is homogeneous if it is a sum of monomials of the same weight.

A vector field f is called *homogeneous of weight j* if for each homogeneous polynomial φ , $f \cdot \varphi$ is a homogeneous polynomial of weight $\mathcal{W}(\varphi) - j$ (the polynomials of negative weights are understood to be equal to 0). Roughly speaking, if the weight of f is j , then f "subtracts" weight j from the functions. In terms of components in the chart \mathbf{x} , f is homogeneous of weight j if and only if its i th component, f^i , is a homogeneous polynomial of weight $w^i - j$.

The *graded order* $\mathcal{O}(\varphi)$ of a function φ is defined as the *minimum weight* of the monomials contained in its Taylor asymptotic expansion at ξ_0 in the chart \mathbf{x} .

The *graded order* $\mathcal{O}(f)$ of a vector field f is defined as the *maximum weight* of the homogeneous vector fields appearing in its Taylor asymptotic expansion at ξ_0 in the chart \mathbf{x} . In other words $\mathcal{O}(f) = j$ if and only if $\mathcal{O}(f^i) \equiv w^i - j$, $i = 1, \dots, n$, with equality for some i .

The graded structure on M at ξ_0 can be extended to the manifold M^* on which the pullback system Σ_{r^*} is defined; see § 3. Without loss of generality we can suppose that $M^* = I \times U$. Denoting by $x^0: M^* \rightarrow \mathbb{R}$ the first canonical projection, we set $\mathbf{x}^* = (x^0, \mathbf{x})$ and $\mathbf{w}^* = (w^0, \mathbf{w})$, where $w^0 \equiv \mathcal{O}(f_0)$. In the chart \mathbf{x}^* the vector field $\partial/\partial t$ is $\partial/\partial x^0$ and $(\mathbf{x}^*, M^*, \mathbf{w}^*)$ defines a local graded structure on M^* at ξ_0 . By noting that

$$f^*(x^0, \xi) = \sum_{i \geq 0} \frac{(x^0)^i}{i!} \text{ad}_{f_0}^i f(\xi),$$

it is easy to see that for each $f \in \mathcal{V}(M)$, $\mathcal{O}(f^*) = \mathcal{O}(f)$, where \mathcal{O} denotes both the graded orders induced by the above graded structures on M and M^* . Moreover, f is homogeneous of weight r if and only if f^* is also (see [3, § 4]).

The systems defined by homogeneous vector fields of positive weight will be called *positively homogeneous systems*. Notice that the definition of local graded structure implies that it makes sense to consider homogeneous systems only if we restrict the system to U . To be more precise, let f_i be homogeneous of weight $l_i > 0$, $i = 0, \dots, m$. In this case, f_i^j is a homogeneous polynomial of weight $w^j - l_i$, so that it does not depend on the x^k 's with $w^k \equiv w^j$. In other words, the system Σ_r is a cascade of integrators. Since the Lie product of two homogeneous vector fields is homogeneous and its weight is the sum of the weights of the two factors, then \mathcal{S}_r is a nilpotent Lie algebra spanned by homogeneous vector fields. Moreover, the f_i 's in the chart (\mathbf{x}, U) are polynomial, so that $N_r(\xi_0, t)$ is an integral submanifold of \mathcal{S}_r . Other properties of positively homogeneous systems can be analyzed by means of a one-parameter family of variations of the null control associated with the set of integers $\mathbf{l} = (l_0, \dots, l_m)$.

Namely, the continuous map

$$\delta : [-1, 1] \times L^1([0, T], \mathbb{R}^m) \rightarrow L^1(\mathbb{R}, \mathbb{R}^m)$$

defined by

$$\delta(\varepsilon, u) \equiv \delta_\varepsilon u \equiv u_\varepsilon : t \mapsto \begin{cases} (\varepsilon^{l_1-l_0} u_1(t/\varepsilon^{l_0}), \dots, \varepsilon^{l_m-l_0} u_m(t/\varepsilon^{l_0})) & \text{if } t \in [0, \varepsilon^{l_0} T] \\ 0 & \text{otherwise.} \end{cases},$$

($\delta(0, u)$ is understood to be identically equal to 0.) The definition of $\delta_\varepsilon u$ implies that if $\alpha = \min \{l_i, i = 1, \dots, m\}$, then

$$(4.1) \quad \|\delta_\varepsilon u\|_{L^1} = \int_0^T \sum_{i=1}^m \varepsilon^{l_i} |u_i(s)| ds \leq \varepsilon^\alpha \|u\|_{L^1}.$$

Moreover, standard calculations imply that $\delta : [-1, 1] \times \mathcal{U} \rightarrow \mathcal{U}$ is continuous. Note that δ_ε maps \mathcal{U} into \mathcal{U} if and only if $\mathbf{l} = (l_0, \dots, l_m)$ is a set of admissible weight for Ω , see Definition 3.4.

The map δ has the following property (see [3, Thm. 2.1b]): if $\Sigma_\mathbf{f}$ is a homogeneous system and $S_\mathbf{f}(t, \xi, u)$ is defined, then

$$(4.2) \quad S_\mathbf{f}(\varepsilon^{l_0} t, \delta_\varepsilon \xi, u_\varepsilon) = \delta_\varepsilon S_\mathbf{f}(t, \xi, u).$$

Therefore, $\forall t \in \mathbb{R}, \forall \omega \in \Omega$, $\delta_\varepsilon \exp t(f_0 + \sum_{i=1}^m \omega_i f_i) \cdot \xi = \exp \varepsilon^{l_0} t(f_0 + \sum_{i=1}^m \varepsilon^{l_i-l_0} \omega_i f_i) \cdot \delta_\varepsilon \xi$. This implies that δ_ε maps $N_\mathbf{f}(\xi_0, t)$ in $N_\mathbf{f}(\xi_0, \varepsilon^{l_0} t)$. The two manifolds have the same dimension and δ_ε is an open map; hence $\delta_\varepsilon \text{int}_{\text{rel}} R_\mathbf{f}(\xi_0, t) \subseteq \text{int}_{\text{rel}} R_\mathbf{f}(\xi_0, \varepsilon^{l_0} t)$. Therefore, if \mathbf{l} is a set of admissible positive weights, then there is t such that

$$\hat{x}_\mathbf{f}(t) \in \text{int}_{\text{rel}} R_\mathbf{f}(\xi_0, t)$$

if and only if $(\Sigma_\mathbf{f}, \Omega)$ is weakly locally controllable along $\hat{x}_\mathbf{f}$. In fact, if such a t exists, then $\hat{x}_\mathbf{f}(\varepsilon^{l_0} t) = \delta_\varepsilon \hat{x}_\mathbf{f}(t) \in \delta_\varepsilon \text{int}_{\text{rel}} R_\mathbf{f}(\xi_0, t) \subseteq \text{int}_{\text{rel}} R_\mathbf{f}(\xi_0, \varepsilon^{l_0} t)$.

For the positively homogeneous systems, the following inverse of Theorem 1.10 holds.

THEOREM 4.1. *Let us suppose that (i) f_i is homogeneous of weight $l_i > 0$, $i = 0, \dots, m$, (ii) $\mathbf{l} = (l_0, \dots, l_m)$ is a set of admissible weights, and (iii) there is $t > 0$ such that $\hat{x}_\mathbf{f}(t) \in \text{int}_{\text{rel}} R_\mathbf{f}(\xi_0, t)$. If $h \in \mathcal{S}_\mathbf{f}$ is a homogeneous vector field of weight r such that $h(\xi_0) \neq 0$, then $h(\xi_0)$ is a regular variation at 0 of order r/l_0 .*

Proof. $h \in \mathcal{S}_\mathbf{f}$ is homogeneous of weight r if and only if $h^* \in \mathcal{S}_{\mathbf{f}^*}$ is and $h(\xi_0) = h^*(\xi_0)$. Therefore, we can prove the theorem for the pullback system. Let h_1, \dots, h_s be homogeneous vector fields that span $\mathcal{S}_{\mathbf{f}^*}$ at ξ_0 , and let r_1, \dots, r_s be their weights. The local inverse of the map

$$(d_1, \dots, d_s) \mapsto \exp d_1 h_1 \cdots \exp d_s h_s \cdot \xi_0$$

defines a chart (\mathbf{y}, W) of $N_{\mathbf{f}^*}(\xi_0, 0)$ at ξ_0 . Set $\mathbf{y}(W) = V$. Lemma A in the appendix of [3] implies, possibly restricting W , that there is a k -tuple $(\omega^1, \dots, \omega^k) \subseteq \Omega$ and an analytic map $\tau : V \rightarrow \{(t_1, \dots, t_k) : t_i \geq 0, t_1 + \dots + t_k = t\}$ such that

$$S_{\mathbf{f}^*}(t, \xi_0, u(\mathbf{d})) = (t, \mathbf{y}^{-1}(\mathbf{d})),$$

where $u(\mathbf{d})$ is the piecewise-constant control with values $\omega^1, \dots, \omega^k$ and switching times $\tau_1(\mathbf{d}), \dots, \tau_k(\mathbf{d})$. Denoting the canonical projection by $p : M^* \rightarrow U$, we obtain

$$\mathbf{y}(p(S_{\mathbf{f}^*}(t, \xi_0, u(\mathbf{d})))) = \mathbf{d}.$$

By (4.2) it follows that

$$\delta_\varepsilon^{-1} S_{\mathbf{f}^*}(\varepsilon^l t, (\varepsilon^l \gamma, \xi_0), \delta_\varepsilon u(\mathbf{d})) = S_{\mathbf{f}^*}(t, (\gamma, \xi_0), u(\mathbf{d})).$$

As a consequence, the map

$$(\gamma, \varepsilon, \mathbf{d}) \mapsto (\gamma, \varepsilon, \mathbf{y}(p(\delta_\varepsilon^{-1} S_{\mathbf{f}^*}(\varepsilon^l t, (\varepsilon^l \gamma, \xi_0), \delta_\varepsilon u(\mathbf{d}))))))$$

is defined in a neighborhood of $(0, 0, 0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^s$, it is analytic, and it has maximum rank at $(0, 0, 0)$. Therefore, it has a local analytic inverse (id, id, ν) . Set $\mathbf{e}_i = (1, 0, \dots, 0)$, and define $\eta(c, \gamma, \varepsilon) = \delta_\varepsilon u(\nu(\gamma, \varepsilon, c\mathbf{e}_i))$. The control η is a continuous function of its arguments, and we obtain

$$\begin{aligned} S_{\mathbf{f}^*}(\varepsilon^l t, \hat{\mathbf{x}}_{\mathbf{f}^*}(\varepsilon^l \gamma), \eta(c, \gamma, \varepsilon)) &= \delta_\varepsilon(t + \gamma, \mathbf{y}^{-1}(c\mathbf{e}_i)) \\ &= (\varepsilon^l(t + \gamma), \delta_\varepsilon \exp ch_1 \cdot \xi_0) = (\varepsilon^l(t + \gamma), \exp \varepsilon^{r_1} ch_1 \cdot \xi_0) \\ &= \hat{\mathbf{x}}_{\mathbf{f}^*}(\varepsilon^l(t + \gamma)) + \varepsilon^{r_1} ch_1(\xi_0) + o(\varepsilon^{r_1+a}) \quad \forall a < r_1. \end{aligned}$$

Setting $\varepsilon' = \varepsilon^l t$, we see that $h_1(\xi_0)$ is a regular variation of order r_1/l_0 . The statement follows because each homogeneous vector field that does not vanish at ξ_0 can be chosen as h_1 . \square

COROLLARY 4.2. *A positively homogeneous system $\Sigma_{\mathbf{f}}$ is weakly locally controllable along the trajectory relative to the null control if and only if $\mathcal{H}_{\mathbf{f}}(0) = \Delta_{\mathbf{f}}^0(\xi_0)$.*

We want to use graded structures at $\hat{\mathbf{x}}_{\mathbf{f}}(t)$ to construct variations at t of the reference couple. Without loss of generality we can suppose $t=0$. The vector fields $f_i \in \mathcal{V}(M)$ can be approximated at ξ_0 by means of vector fields homogeneous with respect to a graded structure $(\mathbf{x}, U, \mathbf{w})$ in the following sense (see [3]). Let $\mathcal{O}(f_i)$ be the graded order of the vector field f_i with respect to $(\mathbf{x}, U, \mathbf{w})$, and let $\mathbf{l} = (l_0, \dots, l_m)$ be a set of positive integers such that $l_i \geq \mathcal{O}(f_i)$. The graded approximation of weight l_i of f_i , say, \hat{f}_i , is the sum of the homogeneous vector fields of weight greater than or equal to l_i in the Taylor asymptotic expansion of f_i at ξ_0 in the chart \mathbf{x} . \hat{f}_i is homogeneous of weight l_i , $i = 0, \dots, m$, and $\hat{f}_i = 0$ if and only if $l_i > \mathcal{O}(f_i)$. Set $\hat{\mathbf{f}} = \{\hat{f}_0, \dots, \hat{f}_m\}$.

DEFINITION 4.3. The system $\Sigma_{\hat{\mathbf{f}}}$ defined on U is called a *graded approximation* at ξ_0 of the system $\Sigma_{\mathbf{f}}$. It is the graded approximation relative to $(\mathbf{x}, U, \mathbf{w})$ induced by \mathbf{l} .

The solutions of a system $\Sigma_{\mathbf{f}}$ and the ones of its graded approximation $\Sigma_{\hat{\mathbf{f}}}$ are related by the dilations of both state and control defined previously (see [3, Thm. 2.1]). In the rest of this paper we will also use the following result:

PROPOSITION 4.4. *Let $\bar{t} \in [0, T]$, $\bar{\xi} \in U$, $\omega = (\omega_1, \dots, \omega_m) \in \Omega$ be such that if $u \equiv \omega$, then $S_{\mathbf{f}}(\cdot, \bar{\xi}, u)$ is defined in $[0, \bar{t}]$. There exists a neighborhood $V(\bar{\xi})$ of $\bar{\xi}$ in U , a neighborhood $V(\bar{t})$ of \bar{t} in \mathbb{R} , an $\bar{\varepsilon} > 0$ such that $\forall \xi \in V(\bar{\xi}), \forall t \in V(\bar{t})$, and $\forall \varepsilon \in (-\bar{\varepsilon}, \bar{\varepsilon})$, $\varepsilon \neq 0$, $\delta_\varepsilon^{-1} S_{\mathbf{f}}(\cdot, \delta_\varepsilon \xi, u_\varepsilon)$ is defined in $[0, \varepsilon^l t]$ (on $[\varepsilon^l t, 0]$ if $\varepsilon^l < 0$). Moreover, the map $H: V(\bar{\xi}) \times V(\bar{t}) \times (-\bar{\varepsilon}, \bar{\varepsilon}) \rightarrow U$ defined by*

$$H(\xi, t, \varepsilon) = \begin{cases} \delta_\varepsilon^{-1} S_{\mathbf{f}}(\varepsilon^l t, \delta_\varepsilon \xi, u_\varepsilon), & \varepsilon \neq 0, \\ S_{\hat{\mathbf{f}}}(t, \xi, u), & \varepsilon = 0, \end{cases}$$

is a C^∞ map.

Proof. It is not restrictive to assume that U is an open neighborhood of 0 in \mathbb{R}^n . Let $\varepsilon \neq 0$, $\xi \in U$; the map $t \mapsto \delta_\varepsilon^{-1} S_{\mathbf{f}}(\varepsilon^l t, \delta_\varepsilon \xi, u_\varepsilon)$ is the solution of the differential equation on U

$$\dot{x} = F_0(\varepsilon, x) + \sum_{i=1}^m \omega_i F_i(\varepsilon, x), \quad x(0) = \xi,$$

where the F_i 's are defined by

$$F_i(\varepsilon, x) = \begin{cases} \varepsilon^{l_i} (\delta_\varepsilon^{-1})_* f_i(\delta_\varepsilon x), & \varepsilon \neq 0, \\ \hat{f}_i(x), & \varepsilon = 0. \end{cases}$$

The F_i 's are C^∞ vector fields on $(-1, 1) \times U$ (see [3, property P5]). The statement follows by the properties of differential equations. \square

COROLLARY 4.5. *Let $u = u(\tau_1, \dots, \tau_k)$ be the piecewise-constant control with values $\omega^1, \dots, \omega^k$ and switching times τ_1, \dots, τ_k . Let $t(\tau) = \sum_{i=1}^k \tau_i$; the map*

$$(\varepsilon, \xi, \tau_1, \dots, \tau_k) \mapsto \begin{cases} \delta_\varepsilon^{-1} S_f(\varepsilon^{l_0} t(\tau), \delta_\varepsilon \xi, u_\varepsilon), & \varepsilon \neq 0 \\ S_f(t, \xi u), & \varepsilon = 0 \end{cases}$$

is C^∞ .

Under suitable assumptions the local controllability of an approximating system along \hat{x}_f implies that all the vector fields in \mathcal{S}_f give rise to variations.

THEOREM 4.6. *Let us suppose that (i) $\mathbf{l} = (l_0, \dots, l_m)$ is a set of admissible positive weights such that $l_i \geq \mathcal{O}(f_i)$, $i = 0, \dots, m$, (ii) $\dim N_f(\xi_0, \cdot) = \dim N_f(\xi_0, 0) = s$, and (iii) there is $t > 0$ such that $\hat{x}_f(t) \in \text{int}_{\text{rel}} R_f(\xi_0, t)$. If $\chi \in \mathcal{S}$ is a bracket of \mathbf{l} -weight r such that $\chi_f(\xi_0) \neq 0$, then $\chi_f(\xi_0)$ is a regular variation at 0 of order r/l_0 for the reference couple of the original system.*

Proof. It is easily seen that starting from a system we obtain the same result if either (a) we first consider its graded approximation relative to $(\mathbf{x}, U, \mathbf{w})$ and \mathbf{l} and then the pullback or (b) first we consider the pullback and then the graded approximation relative to $(\mathbf{x}^*, M^*, \mathbf{w}^*)$ and \mathbf{l} . Therefore, we can prove the theorem for the pullback system. If $f \in \mathcal{V}(M)$, \hat{f}^* will denote the graded approximation of f^* that is equal to $(\hat{f})^*$. Assumption (ii) and [3, Lemma 2.1] imply that there are $h_1, \dots, h_s \in \mathcal{S}_{f^*}$ such that (a) $h_1 = \chi_{f^*}$ and $\text{span}\{h_1(\xi_0), \dots, h_s(\xi_0)\} = \mathcal{S}_{f^*}(\xi_0)$ and (b) $\text{span}\{\hat{h}_1(\xi_0), \dots, \hat{h}_s(\xi_0)\} = \mathcal{S}_{f^*}(\xi_0)$. By possibly restricting U it is also possible to assume that $N_{f^*}(\xi_0, 0)$ (which by (ii) is an integral manifold of \mathcal{S}_{f^*}) is embedded in M^* . Let r_i be the weight of h_i , and define the map $G: (-1, 1) \times (-1, 1)^s \rightarrow \mathbb{R} \times M^*$ locally by

$$(\varepsilon, d_1, \dots, d_s) \mapsto \begin{cases} \delta_\varepsilon^{-1} \exp \varepsilon^{r_1} d_1 h_1 \cdots \exp \varepsilon^{r_s} d_s h_s \cdot \xi_0, & \varepsilon \neq 0, \\ \exp d_1 \hat{h}_1 \cdots \exp d_s \hat{h}_s \cdot \xi_0, & \varepsilon = 0. \end{cases}$$

Corollary 4.5 implies that G is C^∞ . Moreover, $\tilde{G}: (\varepsilon, \mathbf{d}) \mapsto (\varepsilon, G(\varepsilon, \mathbf{d}))$ has maximum rank at $(0, 0)$; therefore, there is a neighborhood \mathcal{W} of $(0, 0)$ in $\mathbb{R} \times \mathbb{R}^s$ such that \tilde{G} is a C^∞ homeomorphism of \mathcal{W} onto $\tilde{G}(\mathcal{W})$. Denote by G_ε the restriction of G to $\mathcal{W}_\varepsilon \equiv \mathcal{W} \cap (\{\varepsilon\} \times \mathbb{R}^s)$. It is clear that $G_\varepsilon(\mathcal{W}_\varepsilon)$ is contained in $\delta_\varepsilon^{-1} N_{f^*}(\xi_0, 0) \subseteq \delta_\varepsilon^{-1} U$ (recall that δ_ε^{-1} is defined only on $\delta_\varepsilon U$). We claim that \mathcal{W} can be chosen so that there is a neighborhood U' of ξ_0 in U such that

$$(4.3) \quad G_\varepsilon(\mathcal{W}_\varepsilon) = \delta_\varepsilon^{-1} N_{f^*}(\xi_0, 0) \cap U'.$$

To prove this, notice that by [3, Lemma 2.1] it is possible to choose homogeneous vector fields k_{s+1}, \dots, k_n , of weights r_{s+1}, \dots, r_n such that

$$\begin{aligned} & \text{span}\{\hat{h}_1(\xi_0), \dots, \hat{h}_s(\xi_0), k_{s+1}(\xi_0), \dots, k_n(\xi_0)\} \\ &= \text{span}\{h_1(\xi_0), \dots, h_s(\xi_0), k_{s+1}(\xi_0), \dots, k_n(\xi_0)\} = T_{\xi_0} U. \end{aligned}$$

Moreover, by possibly restricting U the k_i 's can be chosen in such a way that

$$\exp d_n k_n \cdots \exp d_{s+1} k_{s+1} \cdot \xi \notin N_{f^*}(\xi_0, 0) \quad \text{if } \xi \in N_{f^*}(\xi_0, 0) \text{ and } |d_i| < 1.$$

Define $\Phi: (-1, 1) \times (-1, 1)^s \times (-1, 1)^{n-s} \rightarrow (-1, 1) \times U$ locally by

$$(\varepsilon, d_1, \dots, d_n) \mapsto (\varepsilon, \exp d_n k_n \cdots \exp d_{s+1} k_{s+1} \cdot G_\varepsilon(d_1, \dots, d_s)).$$

Φ is C^∞ , and it has maximum rank at $(0, 0, 0)$, so that there is a neighborhood \mathcal{V} of $(0, 0, 0)$ in $\mathbb{R} \times \mathbb{R}^s \times \mathbb{R}^{n-s}$, $\bar{\varepsilon} > 0$ and a neighborhood U' of ξ_0 in U such that Φ is a homeomorphism of \mathcal{V} onto $(-\bar{\varepsilon}, \bar{\varepsilon}) \times U'$. Since the k_i 's are homogeneous vector fields, we have that

$$\delta_\varepsilon U' = \{\exp \varepsilon^{r_n} d_n k_n \cdots \exp \varepsilon^{r_{s+1}} d_{s+1} k_{s+1} \cdot \delta_\varepsilon G_\varepsilon(d_1, \dots, d_s): (\varepsilon, \mathbf{d}) \in \mathcal{V}\}.$$

Therefore,

$$\delta_\varepsilon U' \cap N_{\mathbf{f}^*}(\xi_0, 0) = \{\delta_\varepsilon G_\varepsilon(d_1, \dots, d_s): (\varepsilon, \mathbf{d}) \in \mathcal{V}\}$$

and

$$U' \cap \delta_\varepsilon^{-1} N_{\mathbf{f}^*}(\xi_0, 0) = U' \cap \delta_\varepsilon^{-1} (N_{\mathbf{f}^*}(\xi_0, 0) \cap \delta_\varepsilon U') = \{G_\varepsilon(d_1, \dots, d_s): (\varepsilon, \mathbf{d}) \in \mathcal{V}\}.$$

If we set

$$\mathcal{W} = \mathcal{V} \cap (\mathbb{R} \times \mathbb{R}^s \times \{0\}),$$

we obtain (4.3).

Lemma A in the appendix of [3] implies, possibly restricting \mathcal{W}_0 , that there is a k -tuple $(\omega^1, \dots, \omega^k) \subseteq \Omega$ and an analytic map $\tau: \mathcal{W}_0 \rightarrow \{(t_1, \dots, t_k): t_i \geq 0, t_1 + \dots + t_k = t\}$ such that

$$S_{\mathbf{f}^*}(t, \xi_0, u(\mathbf{d})) = (t, G_0(\mathbf{d})),$$

where $u(\mathbf{d})$ is the piecewise-constant control with values $\omega^1, \dots, \omega^k$ and switching times $\tau_1(\mathbf{d}), \dots, \tau_k(\mathbf{d})$. Denoting the canonical projection by $p: M^* \rightarrow U$, we obtain

$$G_0^{-1}(p(S_{\mathbf{f}^*}(t, \xi_0, u(\mathbf{d})))) = \mathbf{d}.$$

By Corollary 4.5 the map

$$H: (\gamma, \varepsilon, \mathbf{d}) \mapsto \begin{cases} p(\delta_\varepsilon^{-1} S_{\mathbf{f}^*}(\varepsilon^{l_0} t, (\varepsilon^{l_0} \gamma, \xi_0), \delta_\varepsilon u(\mathbf{d}))), & \varepsilon \neq 0, \\ p(S_{\mathbf{f}^*}(t, \xi_0, u(\mathbf{d}))), & \varepsilon = 0, \end{cases}$$

is defined in a neighborhood \mathcal{W}' of $(0, 0, 0)$ and it is C^∞ . Possibly restricting \mathcal{W}' , $H(\mathcal{W}')$ is contained in U' , so that the map

$$(\gamma, \varepsilon, \mathbf{d}) \mapsto \begin{cases} (\gamma, \varepsilon, G_\varepsilon^{-1}(p(\delta_\varepsilon^{-1} S_{\mathbf{f}^*}(\varepsilon^{l_0} t, (\varepsilon^{l_0} \gamma, \xi_0), \delta_\varepsilon u(\mathbf{d}))))), & \varepsilon \neq 0, \\ (\gamma, 0, G_0^{-1}(p(S_{\mathbf{f}^*}(t, (\gamma, \xi_0), u(\mathbf{d}))))), & \varepsilon = 0, \end{cases}$$

is defined in a neighborhood of $(0, 0, 0)$, it is C^∞ , and its derivative has maximum rank at $(0, 0, 0)$. Therefore, it has a local analytic inverse (id, id, ν) . Set $\mathbf{e}_1 = (1, 0, \dots, 0)$, and define $\eta(c, \gamma, \varepsilon) = \delta_\varepsilon u(\nu(\gamma, \varepsilon, c\mathbf{e}_1))$. The control η is a continuous function of its arguments and so we obtain

$$\begin{aligned} S_{\mathbf{f}^*}(\varepsilon^{l_0} t, (\varepsilon^{l_0} \gamma, \xi_0), \eta(c, \gamma, \varepsilon)) &= (\varepsilon^{l_0}(t + \gamma), \exp \varepsilon^{r_1} c h_1 \cdot \xi_0) \\ &= (\varepsilon^{l_0}(t + \gamma), \xi_0) + \varepsilon^{r_1} c \chi_{\mathbf{f}^*}(\xi_0) + o(\varepsilon^{r_1+a}) \quad \forall a < r_1. \end{aligned}$$

Setting $\varepsilon' = \varepsilon^{l_0} t$, we obtain that $\chi_{\mathbf{f}^*}(\xi_0)$ is a regular variation of order r_1/l_0 . \square

Up to now we have used a graded structure on M^* induced by the graded structure $(\mathbf{x}, U, \mathbf{w})$ on M . This choice of graded structure on M^* allows us to deduce the properties of the graded approximation of the pullback system starting from the properties of the original system. On the other hand, one can choose any graded structure on M^*

and consider directly the properties of the graded approximation of Σ_{r^*} induced by this graded structure. Notice that in this case the properties of $\Sigma_{\hat{r}^*}$ may have no links with the properties of $\Sigma_{\hat{r}}$. Nevertheless, additional results may be obtained by this method. For example, the following result holds true.

COROLLARY 4.7. *Let $(\mathbf{x}^*, M^*, \mathbf{w}^*)$ be any graded structure on M^* at ξ_0 , and let $\Sigma_{\hat{r}^*}$ be the graded approximation of Σ_{r^*} induced by a set of admissible weights $\mathbf{l} = (l_0, \dots, l_m)$ such that $l_0 \geq w^0$, $l_i \geq \mathcal{O}(f_i^*)$, $i = 1, \dots, m$. If (i) $\dim N_{r^*}(\xi_0, 0) = \dim N_{\hat{r}^*}(\xi_0, 0)$, (ii) there is $t > 0$ such that $\hat{x}_{\hat{r}^*}(t) \in \text{int}_{\text{rel}} R_{\hat{r}^*}(\xi_0, t)$, and (iii) $\chi \in \mathcal{S}$ is a bracket of \mathbf{l} -weight r such that $\chi_{\hat{r}^*}(\xi_0) \neq 0$, then $\chi_{\hat{r}}(\xi_0)$ is a regular variation at 0 of order r/l_0 for the reference couple of the original system.*

Proof. By Theorem 4.6 $\chi_{r^*}(\xi_0)$ is a variation of Σ_{r^*} . The statement follows if we take into account that the variations of $\Sigma_{\hat{r}}$ and of Σ_{r^*} coincide and that $\chi_{r^*}(\xi_0) = \chi_{\hat{r}}(\xi_0)$. \square

Up to now we have constructed variations under the hypothesis that an approximating system is locally controllable along its reference trajectory. On the other hand, $\Sigma_{\hat{r}^*}$ is a cascade of integrators, so that some subsystem can be locally controllable along its reference trajectory. If this is the case, we obtain a subspace of $\mathcal{S}_{\hat{r}}(\xi_0)$ as a space of variations. To this aim we choose a suitable graded structure on M^* and we consider suitable subsystems of $\Sigma_{\hat{r}^*}$.

Let $(\mathbf{x}, U, \mathbf{w})$ be a graded structure on M at ξ_0 . Define on M^* the graded structure $\mathbf{w}^* = (w^0, \mathbf{w})$, where w^0 is any positive integer. Let $\mathbf{l} = (l_0, \dots, l_m)$ be a set of positive integers such that $l_0 = w^0$ and $l_i \geq \mathcal{O}(f_i^*)$, $i = 1, \dots, m$. As above, we can consider the graded approximation $\hat{\mathbf{f}}^* = \{\partial/\partial x^0, \hat{f}_1^*, \dots, \hat{f}_m^*\}$ of the family \mathbf{f}^* induced by \mathbf{l} and \mathbf{w}^* .

Let s be an integer less than or equal to $\max\{w^1, \dots, w^n\}$, and let the manifold $N \subseteq U$ be defined by

$$\{\xi \in U: x^j(\xi) = 0 \text{ if } w^j > s\}.$$

Let $N^* = I \times N$, and let $i: N^* \rightarrow M^*$ be the injection map. The chart \mathbf{x}^* induces a projection $\pi: M^* \rightarrow N^*$. For each $f^* \in \mathcal{V}(M^*)$, ϕ will denote the vector field on N^* given by $\phi = \pi_*(f^* \circ i)$, that is, the projection on TN^* of f^* restricted to N^* .

Let Σ_ϕ be the control system associated with the family $\Phi = \{\partial/\partial x^0, \phi_1, \dots, \phi_m\}$, and let $\Sigma_{\hat{\phi}}$ be the approximating system defined by $(\mathbf{x}^*, M^*, \mathbf{w}^*)$ and \mathbf{l} . The main result in this section is that if for some choice of the positive numbers s and w^0 the system $(\Sigma_{\hat{\phi}}, \Omega)$ is weakly locally controllable along $t \mapsto (t, \xi_0)$, then $T_{\xi_0}N$ is a subspace of $\mathcal{H}_{\hat{r}}(0)$. More precisely the following result holds.

THEOREM 4.8. *Let us suppose that (i) $\mathbf{l} = (l_0, \dots, l_m)$ is a set of admissible positive weights such that $l_0 = w^0$ and $l_i \geq \mathcal{O}(\phi_i)$, $i = 1, \dots, m$ and (ii) there exists $t > 0$ such that (t, ξ_0) is interior to $R_{\hat{\phi}}(\xi_0, t)$ relative to $\{t\} \times N$. If $\chi \in \mathcal{S}$ is a bracket of \mathbf{l} -weight r such that $\chi_{\hat{\phi}}(\xi_0) \neq 0$, then $\chi_{\hat{\phi}}(\xi_0)$ is a regular variation at 0 of order r/l_0 for the reference couple of the original system.*

Proof. Applying Theorem 4.6 to the system Σ_ϕ , we obtain that $\chi_\phi(\xi_0)$ is a regular variation of order r/l_0 for $(\hat{x}_\phi, 0)$. Let $\eta(c, \gamma, \varepsilon) = \delta_\varepsilon u(\gamma, \varepsilon, c)$ be the control constructed in that theorem with the property

$$(4.4) \quad S_\phi(\varepsilon^b t, (\varepsilon^b \gamma, \xi_0), \eta(c, \gamma, \varepsilon)) = (\varepsilon^b(\gamma + t), \varepsilon^r \chi_\phi(\xi_0) + o(\varepsilon^{r+a})).$$

Since $\Sigma_{\hat{r}^*}$ is a cascade of integrators and $\|u(\gamma, \varepsilon, c)\|_{L^1}$ is uniformly bounded, there is $\sigma > 0$ such that $S_{\hat{r}^*}(\sigma^b \tau, (\sigma^b \gamma, \xi_0), \delta_\sigma u(\gamma, \varepsilon, c))$ is defined and belongs to M^* for $\tau \in [0, t]$. Therefore, [3, Thm. 2.1] applies, so that we obtain that $S_{r^*}(\varepsilon^b t, (\varepsilon^b \gamma, \xi_0), \eta(c, \gamma, \varepsilon))$ is defined for ε smaller than σ . Moreover, for each j we have

$$(4.5) \quad \varepsilon^{-w^j} x^j(S_{r^*}(\varepsilon^b t, (\varepsilon^b \gamma, \xi_0), \eta(c, \gamma, \varepsilon))) \quad \text{uniformly bounded.}$$

Set $u_\varepsilon = \eta(c, \gamma, \varepsilon)$; to end the proof of the statement it is sufficient to show that

$$\|S_{f^*}(\varepsilon^b t, (\varepsilon^b \gamma, 0), u_\varepsilon) - S_\Phi(\varepsilon^b t, (\varepsilon^b \gamma, 0), u_\varepsilon)\| = o(\varepsilon^{s+\alpha}) \quad \forall \alpha < 1$$

uniformly on γ and c . By the definition and by (4.5) we have

$$\begin{aligned} & \|S_{f^*}(\varepsilon^b t, (\varepsilon^b \gamma, 0), u_\varepsilon) - S_\Phi(\varepsilon^b t, (\varepsilon^b \gamma, 0), u_\varepsilon)\| \\ & \leq \|\pi S_{f^*}(\varepsilon^b t, (\varepsilon^b \gamma, 0), u_\varepsilon) - S_\Phi(\varepsilon^b t, (\varepsilon^b \gamma, 0), u_\varepsilon)\| + o(\varepsilon^{s+\alpha}) \\ & \leq \int_0^{\varepsilon^b t} \sum_{i=1}^m |(u_\varepsilon(\tau))_i| \|\pi_* f_i^*(S_{f^*}(\tau, (\varepsilon^b \gamma, 0), u_\varepsilon)) \\ & \quad - \pi_* f_i^*(S_\Phi(\tau, (\varepsilon^b \gamma, 0), u_\varepsilon))\| d\tau + o(\varepsilon^{s+\alpha}) \\ & \leq \int_0^{\varepsilon^b t} \sum_{i=1}^m |(u_\varepsilon(\tau))_i| L_i \|S_{f^*}(\tau, (\varepsilon^b \gamma, 0), u_\varepsilon) - S_\Phi(\tau, (\varepsilon^b \gamma, 0), u_\varepsilon)\| d\tau + O(\varepsilon^{s+\alpha}), \end{aligned}$$

where L_i are local Lipschitz constants of the f_i 's. The L_1 -norm of u_ε is uniformly bounded, so that Gronwall's lemma completes the proof. \square

Proof of Theorem 3.5. The proof of the theorem is based on the results in [3]. Notice that the proof given in [3] is for the case in which Ω is a hypercube, i.e., $\rho_1 = \dots = \rho_m$; however, it is easy to see that the proofs do not change if Ω is a polydisk. We can suppose that \mathbf{l} is a set of *positive* weights. In fact, if this is not the case, we can construct a new set of admissible positive weights for which hypothesis (3.5) of the theorem is fulfilled (see [3, Lemma 4.2]). Moreover, a bracket Λ is \mathbf{l} -neutralized for Σ_f if and only if it is \mathbf{l} -neutralized for Σ_{f^*} (see [3, Lemma 4.1]). Therefore, it is sufficient to prove the theorem for the pullback system Σ_{f^*} when \mathbf{l} is a set of *positive* weights. \mathbf{l}^* defines a filtration $\mathcal{N} = \{N_i\}_{i \geq 0}$ of \mathcal{S}_f , where $N_i = \text{span} \{\Lambda_f : \Lambda_f \in \mathcal{S}, \|\Lambda_f\|_1 \leq i\}$. Let $(\mathbf{x}, U, \mathbf{w})$ be a graded structure induced by \mathcal{N} at ξ_0 up to weight $p = \min\{i : N_i(\xi_0) = \mathcal{S}_f(\xi_0)\}$ (see [3, §§ 3, 4]). We set $\mathbf{x}^* = (x^0, \mathbf{x})$ and $\mathbf{w}^* = (l_0, \mathbf{w})$. $(\mathbf{x}^*, M^*, \mathbf{w}^*)$ is such that for each $f \in \mathcal{S}_f$, $\mathcal{O}(f^*) = \mathcal{O}(f)$, where \mathcal{O} denotes both the graded orders induced by the above graded structures on M and M^* (see [3, Lemma 4.3]). We define the submanifold $N \subseteq U$ by $x^j|_N = 0$ if $w^j > s = \|\Phi\|_1$, and we consider the system Σ_Φ on N^* as defined previously. By the definition of the graded structure induced by \mathbf{l}

$$T_{\xi_0} N = \{\Lambda_f(\xi_0) : \|\Lambda_f\|_1 \leq s\}.$$

Moreover, if h is a vector field whose graded order is equal to r , then each component h^j has graded order greater than or equal to $w^j - r$. Therefore, if Λ is a bracket in \mathcal{S} such that $\|\Lambda\|_1 = r$, we obtain (i) if $r \leq s$, then $\Lambda_f(\xi_0) = \Lambda_{f^*}(\xi_0) = \Lambda_\Phi(\xi_0)$, and (ii) if $r > s$, then $\Lambda_\Phi \equiv 0$. If $r \leq s$, [3, Prop. 3.1] implies that $\Lambda_\Phi(\xi_0) = \Lambda_\Phi(\xi_0) \pmod{V_{r-1}^1}$. As a consequence, we obtain

$$\chi_\Phi(\xi_0) = 0 \quad \forall \chi \in \mathcal{B}_1^* \text{ and } \mathcal{S}_\Phi(\xi_0) = T_{\xi_0} N.$$

Applying [3, Thm. 1.1], we obtain that (t, ξ_0) is interior to $R_\Phi(\xi_0, t)$ relative to $\{t\} \times N$ for each t . Theorem 4.8 completes the proof. \square

Remark 4.9. Let $\Phi \in \mathcal{S}$ satisfy the hypotheses of Theorem 3.5 for a set of weights \mathbf{l} that can be chosen to be positive (see [3, Lemma 4.2]). Φ defines a variation of order $\alpha = \|\Phi\|_1/l_0$. The L^1 norm of its control variation η is an $o(\varepsilon^b)$ for each $b < \min\{l_1, \dots, l_m\}/l_0$. Remark 3.2 implies that Φ defines a variation of order α also if the vector fields f_i 's are perturbed with terms of order greater than $k = \|\Phi\|_1/\min\{l_0, \dots, l_m\}$.

The result is not surprising since k is at least as large as the length of each bracket involved in the assumptions.

5. Applications to local controllability along a trajectory. This section is devoted to deriving from Theorem 3.5 sufficient conditions of weak local controllability along a trajectory. We will show that the variational approach we propose unifies most of the known sufficient conditions of local controllability.

By means of Theorem 3.5 we can construct for each admissible set of weights \mathbf{l} a space $V_r^{\mathbf{l}}$ of variations at time 0 as the maximal \mathbf{l} -neutralized subspace of \mathcal{S}_r . If the sum of such spaces over all the set of admissible weights is equal to $\Delta_r^0(\xi_0)$, then by Theorem 1.13 \hat{x}_r is weakly locally controllable. As a consequence, we obtain that different sets of weights can be used to neutralize the obstructions in the sense of the following result.

PROPOSITION 5.1. *Let Ω be a polydisk (possibly unbounded). For each set \mathbf{l} of nonnegative admissible weights let us define $r(\mathbf{l})$ as the maximum integer r for which each $\chi \in \mathcal{B}_1^*$ such that $\|\chi\|_1 \leq r$ is \mathbf{l} -neutralized at ξ_0 . If*

$$\mathcal{W} = \sum_{\mathbf{l}} \text{span} \{h(\xi_0) : h \in V_{r(\mathbf{l})}^{\mathbf{l}}\} = \Delta_r^0(\xi_0),$$

then (Σ_r, Ω) is weakly locally controllable along \hat{x}_r .

Proposition 5.1 contains [3, Thm. 1.1] as a particular case in which only one admissible set of weights is considered for neutralizing the obstructions. We recall that such a result generalizes those in [12] and [18] (see [3, Lemma 1.2] and subsequent paragraphs). The following example shows that more than one set of weights may be necessary to prove the local controllability of a trajectory.

Example 5.2. Let $M = \mathbb{R}^9$, $m = 1$, $\Omega = [-1, 1]$,

$$f_0(x_1, \dots, x_9) = x_1 \frac{\partial}{\partial x_2} + \dots + x_6 \frac{\partial}{\partial x_7} + (x_7 + x_1^4) \frac{\partial}{\partial x_8} + (x_2^2 + x_1^3) \frac{\partial}{\partial x_9},$$

$$f_1(x_1, \dots, x_9) = \frac{\partial}{\partial x_1}.$$

Σ_r is locally controllable at 0. In fact, $\mathbf{l} = (0, 1)$ and $\bar{\mathbf{l}} = (1, 1)$ are sets of admissible weights. $r(\mathbf{l}) = 1$, $r(\bar{\mathbf{l}}) = 4$, and

$$V_1^{\mathbf{l}} = \text{span} \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_8} \right) \quad \text{and} \quad V_4^{\bar{\mathbf{l}}} = \text{span} \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_4}, \frac{\partial}{\partial x_9} \right).$$

The nonzero obstructions are

$$ad_{f_1}^4 f_0 = 4! \frac{\partial}{\partial x_8} \quad \text{and} \quad ad_{[f_1, f_0]}^2 f_0 = 2 \frac{\partial}{\partial x_9}.$$

$ad_{f_1}^4 f_0$ can be neutralized at 0 only by $ad_{f_0}^7 f_1 = \partial / \partial x_8$; $ad_{[f_1, f_0]}^2 f_0$ can be neutralized only by $ad_{f_1}^3 f_0 = 24x_1(\partial / \partial x_8) + 3!(\partial / \partial x_9)$. Suppose that there is a set of weights for which both of the above obstructions are neutralized. In this case, $4l_1 + l_0 > 7l_0 + l_1$ and $2l_1 + 3l_0 > 3l_1 + 3l_0$, i.e., $l_1 > 2l_0$ and $l_1 < 2l_0$, a contradiction.

Remark 5.3. We point out that for applying Proposition 5.1 it is not sufficient that each obstruction be neutralized by a set of admissible weights. For example, consider the system in \mathbb{R}^5 , with $m = 1$, $\Omega = [-1, 1]$,

$$f_0(x_1, \dots, x_5) = x_1 \frac{\partial}{\partial x_2} + (x_2^2 + x_1^3) \frac{\partial}{\partial x_3} + x_3 \frac{\partial}{\partial x_4} + (x_1^4 + x_4) \frac{\partial}{\partial x_5},$$

$$f_1(x_1, \dots, x_5) = \frac{\partial}{\partial x_1}.$$

The nonvanishing obstructions are

$$\Lambda_1 = ad_{[f_0, f_1]}^2 f_0 = 2 \frac{\partial}{\partial x_3} \quad \text{and} \quad \Lambda_2 = ad_{f_1}^4 f_0 = 4! \frac{\partial}{\partial x_5}.$$

They can be neutralized, respectively, by $ad_{f_1}^3 f_0 = 3! \partial / \partial x_3$ and $ad_{f_0}^2 ad_{f_1}^3 f_0 = 3! \partial / \partial x_5$, so that Λ_1 is \mathbf{l} -neutralized if and only if $l_1 < 2l_0$ and Λ_2 is \mathbf{l} -neutralized if and only if $l_1 > 2l_0$. Both Λ_1 and Λ_2 can be neutralized at 0; nevertheless, if $l_1 > 2l_0$, Λ_1 cannot be neutralized and its weight is smaller than that of Λ_2 . Therefore $\partial / \partial x_5$ does not belong to \mathcal{W} and Proposition 5.1 does not apply.

If all the obstructions in $V_s^{\mathbf{l}}$ are neutralized along the trajectory, Theorem 2.4 implies that for each $h \in V_s^{\mathbf{l}}$, $ad_{f_0}^i h(\hat{x}(t))$ is an element of $\mathcal{H}_t(t)$ for each $i \geq 0$ and each $t > 0$. Notice that $ad_{f_0}^i h(\xi_0)$ need not be a variation at 0; nevertheless, it defines directions that can be used to prove controllability. All of the above arguments leads to the following summarizing result.

THEOREM 5.4. *Let Ω be a polydisk (possibly unbounded). For each set \mathbf{l} of nonnegative admissible weights let us define (i) $r(\mathbf{l})$ as the maximum integer r for which each $\chi \in \mathcal{B}_t^*$ such that $\|\chi\|_1 \leq r$ is \mathbf{l} -neutralized at ξ_0 and (ii) $s(\mathbf{l})$ as the maximum integer s for which each $\chi \in \mathcal{B}_s^*$ such that $\|\chi\|_1 \leq s$ is \mathbf{l} -neutralized at $\hat{x}_t(t)$ for all $t \in (0, \tau)$. If*

$$\mathcal{V} = \sum_{\mathbf{l}} \text{span} \{ \{h(\xi_0) : h \in V_{r(\mathbf{l})}^{\mathbf{l}}\} \cup \{ad_{f_0}^i g(\xi_0) : g \in V_{s(\mathbf{l})}^{\mathbf{l}}, i \geq 0\} \} = \Delta_r^0(\xi_0),$$

then (Σ_r, Ω) is weakly locally controllable along \hat{x}_t .

Proof. Lemma 3.1 in [16] states that if each \mathbf{l} -homogeneous element $\chi \in \mathcal{B}_s^{\mathbf{l}} \cap V_s^{\mathbf{l}}$ is \mathbf{l} -neutralized at any point $\hat{x}(t)$ and t belongs to a nondegenerate interval $[0, \tau)$, then each element $\chi \in \mathcal{B}_t^* \cap V_s^{\mathbf{l}}$ is \mathbf{l} -neutralized at the same points. Therefore, each $g \in V_{s(\mathbf{l})}^{\mathbf{l}}$ defines a variation at each $t \in (0, \tau)$. Let

$$\{h_1(\xi_0), \dots, h_k(\xi_0), ad_{f_0}^{i_1} g_1(\xi_0), \dots, ad_{f_0}^{i_j} g_j(\xi_0)\}$$

be a base of $\Delta_r^0(\xi_0)$. By continuity, if t is sufficiently small, the vectors

$$\{(\exp tf_0)_* h_1(\xi_0), \dots, (\exp tf_0)_* h_k(\xi_0), ad_{f_0}^{i_1} g_1(\exp tf_0 \cdot \xi_0), \dots, ad_{f_0}^{i_j} g_j(\exp tf_0 \cdot \xi_0)\}$$

are linearly independent. Since Δ_r^0 has constant dimension on \hat{x}_t , they are a base of $\Delta_r^0(\hat{x}_t(t))$ for t sufficiently small. By the definition of the variational cone and by Theorem 2.4 the vector space they span is contained in $\mathcal{H}_t(t)$. Hence for each $t > 0$ sufficiently small $\mathcal{H}_t(t) = \Delta_r^0(\hat{x}_t(t))$, and Theorem 1.10 completes the proof. \square

Remark 5.5. Theorem 5.4 points out that \hat{x}_t may be weakly locally controllable even if some obstructions are not neutralized by any set of weights. For example, let $M = \mathbb{R}^5$, $m = 1$,

$$f_0(x_1, \dots, x_5) = x_1 \frac{\partial}{\partial x_2} + x_1^3 \frac{\partial}{\partial x_3} + x_3 \frac{\partial}{\partial x_4} + (x_2^2 + x_4) \frac{\partial}{\partial x_5},$$

$$f_1(x_1, \dots, x_5) = \frac{\partial}{\partial x_1}.$$

In this example the obstruction $[[f_0, f_1], [[f_0, f_1], f_0]](0)$ belongs only to

$$\text{span} \{ad_{f_0}^2 [f_1 [f_1 [f_1, f_0]]](0)\},$$

so that it cannot be neutralized by any set of weights. However, setting $\mathbf{l} = (1, 1)$, we obtain $V_{s(\mathbf{l})}^{\mathbf{l}} = \text{span} \{\partial / \partial x_1, \partial / \partial x_2, \partial / \partial x_3\}$ and $\mathcal{V} = \mathbb{R}^5$, so that by Theorem 5.4 Σ_r is locally controllable at 0.

Neutralization with respect to different weights in the case of a stationary trajectory and bounded controls, has been considered by the present authors in [1]. Let us now analyze the case of unbounded controls. Corollary 3.8 implies that if

$$\mathcal{Y}(\xi_0) = \text{span} \{ad_{f_0}^h Y(\xi_0) : Y \in \text{Lie}\{f_1, \dots, f_m\}, h \geq 0\} = \Delta_r^0(\xi_0),$$

then the trajectory \hat{x}_r is weakly locally controllable. Otherwise, it is sufficient to look for a subspace transversal to $\mathcal{Y}(\xi_0)$ in \mathcal{W} . A sufficient condition of weak local controllability involving the brackets that contain f_0 only once is given in [3]. The result is obtained by giving a sufficiently large weight to f_0 and by giving a weight of unity to the controlled vector fields. The same arguments can be used to find subspace of variations. Namely, let

$$W_r = \text{span} \{\chi_r, \chi \text{ contains one } X_0 \text{ and its length is } r+1\}.$$

If s is such that

$$W_{2r}(\xi_0) \subseteq W_{2r-1}(\xi_0) \cup \text{Lie}\{f_1, \dots, f_m\}(\xi_0) \quad \forall 2r \leq s,$$

then $W_s(\xi_0)$ is a subspace of variation at 0. In [9] the author introduces the subspaces $W_r^i = \text{span} \{ad_{f_i}^j f_0, j \leq r\}$ in order to give conditions of local controllability at ξ_0 . The same type of conditions define W_r^i as subspaces of variations. This can be done by grading the weights of the f_i 's in a suitable way and by taking l_0 sufficiently large. Summarizing, we obtain

THEOREM 5.6. *Let $\Omega = \mathbb{R}^m$. Define $J = \text{Lie}(f_1, \dots, f_m)$, $\mathcal{Y} = \text{span} \{ad_{f_0}^h Z, Z \in J\}$, $W_r^i = \text{span} \{ad_{f_i}^j f_0, j \leq r\}$, $s(i)$ the maximum of the integers σ such that $W_{2r}^i(\xi_0) \subseteq W_{2r-1}^i(\xi_0) \cup J(\xi_0) \quad \forall 2r \leq \sigma$, $W_r = \text{span}(\chi_r, \chi \text{ contains one } X_0 \text{ and its length is } r+1)$, and s the maximum integer k with the property that $W_{2r}(\xi_0) \subseteq W_{2r-1}(\xi_0) \cup J(\xi_0) \quad \forall 2r \leq k$. If*

$$\mathcal{Y}(\xi_0) + W_{s(1)}^1(\xi_0) + \dots + W_{s(m)}^m(\xi_0) + W_s(\xi_0) = \Delta_r^0(\xi_0),$$

then (Σ_r, Ω) is weakly locally controllable along \hat{x}_r .

Proof. What we have said previously proves that $\mathcal{Y}(\xi_0) \subseteq \mathcal{W}$. Let $\chi^1, \dots, \chi^r \in \text{Lie}(X_1, \dots, X_m)$ be such that $\chi_r^1(\xi_0), \dots, \chi_r^r(\xi_0)$ is a basis of $J(\xi_0)$, and let h be the maximum length of the χ^i 's. Let $\mathbf{l} = (hs(i), s(i), \dots, 1, \dots, s(i))$ (1 in the i th position). $\Phi_r \in W_{s(i)}^i$ implies $\|\Phi\|_1 \leq (h+1)s(i) = k$. Let $\psi \in \mathcal{B}_1^*$; if ψ contains more than one X_0 , $\|\psi\|_1 > 2hs(i) > k$; if ψ contains any X_j with $j \neq i$, then $\|\psi\|_1 > (h+2)s(i) > k$. Hence the only obstructions with weight less than k belong to $\text{Lie}(X_0, X_i)$, contain X_0 once, and contain X_i no more than $s(i)$ times. The hypotheses imply that these obstructions are neutralized, since $\|\chi^j\|_1 \leq hs(i), j = 1, \dots, r$. Hence $W_{s(i)}^i \subseteq \mathcal{W}$. Let $\mathbf{m} = (hs, 1, \dots, 1)$ (h is the integer defined above). $\Phi_r \in W_s$ implies $\|\Phi\|_m \leq (h+1)s = \bar{k}$. The only obstructions with weight less than \bar{k} contain one X_0 and at most s indeterminates with indices different from 0. By the hypotheses these obstructions are neutralized, since $\|\chi^j\|_m \leq h < hs$. Hence $W_s \subseteq \mathcal{W}$, and the proof is complete. \square

REFERENCES

- [1] R. M. BIANCHINI AND G. STEFANI, *Sufficient conditions of local controllability*, in Proc. 25th IEEE Conference on Decision and Control, 1986, pp. 967-970.
- [2] ———, *A high order maximum principle*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988, pp. 131-136.
- [3] ———, *Graded approximations and controllability along a trajectory*, SIAM J. Control Optim., 28 (1990), pp. 903-924.

- [4] ———, *A note on the reachable sets of control Systems*, in New Trends in System Theory, G. Conte, A. M. Perdon, and B. Wyman, eds., Progress in Systems and Control Theory, 7, Birkhäuser, Boston, 1991, pp. 113–119.
- [5] A. BRESSAN, *A high order test for optimality of bang-bang controls*, SIAM J. Control Optim., 23 (1985), pp. 38–48.
- [6] D. L. ELLIOT AND N. KALOUPSIDIS, *Accessibility properties of smooth nonlinear control system*, in the 1975 Ames Research Center NASA Conference on Geometric Control Theory, Vol. VII, C. Martin and R. Hermann, eds., Mathematics-Science Press, Brooklyn, NY, 1977, pp. 439–446.
- [7] H. FRANKOWSKA, *Local Controllability of control systems with feedback*, J. Optim. Theory Appl., 60 (1989), pp. 277–296.
- [8] ———, *Contingent cones to reachable sets of control systems*, SIAM J. Control Optim., 27 (1989), pp. 170–198.
- [9] J. B. GONCALVES, *Sufficient conditions for local controllability with unbounded controls*, SIAM J. Control Optim., 25 (1987), pp. 1371–1378.
- [10] K. A. GRASSE, *Controllability and accessibility in nonlinear control systems*, PH.D. thesis, University of Illinois, Urbana 1979.
- [11] K. A. GRASSE AND H. J. SUSSMANN, *Global controllability by nice controls*, in Nonlinear Controllability and Optimal Control Monographs and Textbooks in Pure and Applied Math., 133, H. Sussmann, ed., Marcel-Dekker, New York, 1990.
- [12] H. HERMES, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.
- [13] H. W. KNOBLOCH, *Higher order necessary conditions in optimal control theory*, Lecture Notes in Control and Information Sciences 34, Springer-Verlag, Berlin, New York, 1981.
- [14] A. KRENER, *The high order maximal principle and its applications to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–292.
- [15] G. STEFANI, *On local controllability of the scalar input control systems*, in Theory and Applications of Nonlinear Control Systems, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 167–179.
- [16] ———, *On the minimum time problem*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988, pp. 213–220.
- [17] H. J. SUSSMAN, *Lie brackets and local controllability: a sufficient condition for scalar input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.
- [18] ———, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
- [19] ———, *Orbits of families of vector fields and integrability of systems with singularities*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.
- [20] H. J. SUSSMAN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

VERSIONS OF SONTAG'S "INPUT TO STATE STABILITY CONDITION" AND THE GLOBAL STABILIZABILITY PROBLEM*

JOHN TSINIAS†

Abstract. This paper deals with the global feedback stabilizability problem for interconnected nonlinear systems that are affine in the control. The corresponding feedback stabilizers are supposed to be almost smooth real mappings. The results of this paper extend those developed in recent works of the author [*SIAM J. Control Optim.*, 29 (1991), pp. 457-473], [*Systems Control Lett.*, 15 (1990), pp. 441-448], [*SIAM J. Control Optim.*, 30 (1992), pp. 879-893] concerning the local stabilizability problem. The main sufficient conditions presented are of Lyapunov type, and the main idea of the paper is based on the input to state stability condition introduced by Sontag [*IEEE Trans. Automat. Control*, 35 (1990), pp. 473-477].

Key words. global stabilization, input to state stability condition, control Lyapunov function

AMS subject classification. 93D15

1. Introduction. In the last few years, there has been an increasing interest in the problem of local and global stabilization of nonlinear control systems by feedback (see, for instance, [1]-[12], [14]-[18], [20]-[36], and [38]).

In particular, in [8], [18], [21], [25], and [29], it was proved that if a smooth nonlinear system is globally smoothly stabilizable at its equilibrium, then adding an integrator does not change this property.

Generalizations of the previous result are also given in [25], [31], and [33]. Further progress was provided in [18] by Kokotovic and Sussmann, where sufficient conditions are established for global stabilization for interconnected systems of the form

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2) \\ Ax_2 \end{pmatrix} + \begin{pmatrix} 0 \\ Bu \end{pmatrix},$$

where A and B are constant matrices.

Our purpose is to provide sufficient conditions for global stabilization for the case of systems

$$(1.1a) \quad \dot{x} = F(x) + G(x)u, \quad (x, u) \in \mathbb{R}^n \times \mathbb{R}^m,$$

$$(1.1b) \quad F(x) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix}, \quad G(x) = \begin{pmatrix} 0 \\ g(x_1, x_2) \end{pmatrix}, \quad x = (x'_1, x'_2)' \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2},$$

or, equivalently,

$$(1.2) \quad \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} + \begin{pmatrix} 0 \\ g(x_1, x_2)u \end{pmatrix}.$$

The stabilization approach we present is based on a version of the *input to state stability condition* introduced by Sontag in [25] and also on some ideas from our recent works [32], [33].

We assume that the mappings $f_1: \mathbb{R}^n \rightarrow \mathbb{R}^{n_1}$, $f_2: \mathbb{R}^n \rightarrow \mathbb{R}^{n_2}$, and $g: \mathbb{R}^n \rightarrow \mathbb{R}^{n_2}$ are continuous and that the origin $O \in \mathbb{R}^n$ is an equilibrium for the uncontrolled term F (i.e., $F(0) = 0$).

* Received by the editors December 17, 1990; accepted for publication (in revised form) October 11, 1991.

† National Technical University, Department of Mathematics, Zografou Campus, 15773 Athens, Greece.

We say that (1.1a) is *globally stabilizable (at the origin)* (G.S.) if there exists a map $r: \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that, for every initial state $x_0 \in \mathbb{R}^n$, the solution $x(t, x_0)$ of the closed-loop system

$$(1.3) \quad \dot{x} = F(x) + G(x)r(x)$$

starting at x_0 , is defined—not necessarily uniquely—for all $t \geq 0$, tending to zero as $t \rightarrow +\infty$ or reaching zero at a finite time and, furthermore, zero is stable with respect to (1.3).

We say that a real function $\Phi: \mathbb{R}^n \rightarrow \mathbb{R}^+$ is a *global control Lyapunov function* (G.C.L.F.) of the origin $O \in \mathbb{R}^n$ with respect to (1.1a) if it is positive definite (namely, $\Phi(0) = 0$ and $\Phi(x) > 0$ otherwise), is continuous on \mathbb{R}^n and at least continuously differentiable on $\mathbb{R}^n \setminus \{0\}$, and, furthermore, (i) Φ is uniformly unbounded on \mathbb{R}^n (namely, $\Phi(x) \rightarrow +\infty$ as $\|x\| \rightarrow +\infty$); and (ii) the following property is satisfied:

$$(1.4) \quad (D\Phi G)(x) \stackrel{\text{def}}{=} ((D\Phi G_1)(x), \dots, (D\Phi G_m)(x)) = 0, \quad x \neq 0 \Rightarrow (D\Phi F)(x) < 0.$$

Artstein's theorem [3] asserts that system (1.1a) admits a G.C.L.F. at zero if and only if it is G.S. by means of a feedback law r that is smooth on $\mathbb{R}^n \setminus \{0\}$. To be more precise, $O \in \mathbb{R}^n$ will be globally asymptotically stable with respect to the resulting system (1.3), where, for each x_0 , the solution $x(t, x_0)$ of (1.3) starting at x_0 is uniquely defined for all positive t .

Further generalizations of the previous theorem are provided in [2], [12], [24], [29]–[31], [35]. In particular, Sontag in [24] offers an explicit formula for the stabilizing feedback law.

Given a continuous map $\phi: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$, we say that the function $W: \mathbb{R}^n \rightarrow \mathbb{R}^+$ is a G.C.L.F. of the set

$$M_\phi \stackrel{\text{def}}{=} \{x = (x'_1, x'_2)' \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}: x_2 = \phi(x_1)\}$$

with respect to (1.2) if W is continuously differentiable on \mathbb{R}^n ; furthermore,

(i) There exist real functions a_i , $i = 1, 2$, of class \mathcal{K}_∞ (namely, a_i is continuous, strictly increasing and satisfies $a_i(0) = 0$ and $a_i(s) \rightarrow +\infty$ as $s \rightarrow +\infty$) such that

$$(1.5) \quad a_1(\|x_2 - \phi(x_1)\|) \leq W(x) \leq a_2(\|x_2 - \phi(x_1)\|)$$

for all $x = (x'_1, x'_2)' \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$;

(ii) There exists a real function a_3 of class \mathcal{K} (namely, a_3 is continuous, strictly increasing with $a_3(0) = 0$) such that the following holds:

$$(1.6) \quad (DWG)(x) = 0, \quad x \notin M_\phi \Rightarrow (DWF)(x) < -a_3(\|x_2 - \phi(x_1)\|).$$

We say that the function W , above, satisfies the *continuity property* (C.P.) if, furthermore, there exists a map $d: \mathbb{R}^n \rightarrow \mathbb{R}^+$ with $d(x) \rightarrow 0$ as $\text{dist}(x, M_\phi) \rightarrow 0$ such that, for any $x \neq 0$, a vector $u \in \mathbb{R}^m$ can be found satisfying the following inequalities:

$$(DWF)(x) + (DWG)(x)u < -a_3(\|x_2 - \phi(x_1)\|), \quad \|G(x)u\| \leq d(x).$$

Note that the above definition generalizes the notion of the *small control property* (see [3], [24], and [29]).

Finally, we say that the system

$$(1.7) \quad \dot{x}_1 = f(x_1, v), \quad (x_1, v) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$$

satisfies the *input to state stability condition* (I.S.S.C.) if there exist a continuously differentiable Lyapunov function $V: \mathbb{R}^n \rightarrow \mathbb{R}^+$ with respect to $\dot{x}_1 = f(x_1, 0)$ (namely,

$V(0) = 0$; $V(x_1) > 0$ and $DV(x_1)f(x_1, 0) < 0$ for $x_1 \neq 0$), which is uniformly unbounded on \mathbb{R}^n , and positive constants σ and ξ such that the following holds:

$$(1.8) \quad \|x_1\| > \xi, \quad \|v\| < \sigma \Rightarrow DV(x_1)f(x_1, v) < 0,$$

and, furthermore, for any measurable essentially bounded input v , system (1.7) is complete.

Note that the above property weakens a well-known Lyapunov-like condition provided by Sontag, which guarantees input to state stabilization (see condition (16) and Theorem 1 in [25]). We also note that (1.8) is satisfied if, for instance, we assume that zero $O \in \mathbb{R}^{n_1}$ is globally exponentially stable with respect to $\dot{x}_1 = f(x_1, 0)$, and, furthermore, the map $f(x_1, v)$ is Lipschitz continuous with respect to v uniformly on x_1 [37].

Consider now the simplest case of system (1.2), where $m = n_2$ and g is the unit matrix of dimension $m \times m$. Suppose that there exists a continuously differentiable map $\phi: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^m$ such that $O \in \mathbb{R}^{n_1}$ is globally asymptotically stable with respect to

$$(1.9) \quad \dot{x}_1 = f_1(x_1, \phi(x_1))$$

and let $V(x_1)$ be a uniformly unbounded smooth Lyapunov function of $O \in \mathbb{R}^{n_1}$ with respect to (1.9). Then the function $\Phi(x) = V(x_1) + W(x)$, where $W(x) = \frac{1}{2}\|x_2 - \phi(x_1)\|^2$ is a G.C.L.F. of $O \in \mathbb{R}^n$ with respect to (1.2), and so the overall system (1.2) is G.S. by means of a feedback law, which is smooth for $x \neq 0$ and continuous at zero (see [29], [31], and [33]).

Our purpose is to extend the previous result to the general nonlinear case (1.2). In particular, in Theorem 2.2, we establish that if there exists a continuous map $\phi: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$ such that the set M_ϕ is positively invariant with respect to $\dot{x} = F(x)$, the system

$$(1.10) \quad \dot{x}_1 = f_1(x_1, \phi(x_1) + v), \quad (x_1, v) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$$

satisfies the I.S.S.C.; furthermore, if (1.2) admits a G.C.L.F. with respect to M_ϕ , then system (1.2) is G.S. by means of an “almost smooth” feedback law $u = r(x)$. In particular, the map r will be smooth on $\mathbb{R}^n \setminus M_\phi$. In Theorem 2.4, we specialize the previous result to homogeneous nonlinear control systems.

If certain additional assumptions for subsystem (1.10) are imposed, then, in Theorems 3.1 and 3.5, we establish that there exists a G.C.L.F. of zero $O \in \mathbb{R}^n$, and so system (1.2) is G.S. by means of a feedback law that is smooth on $\mathbb{R}^n \setminus \{0\}$. In particular, in Theorem 3.1 we show that, if, in addition to the hypothesis of Theorem 2.2, we assume that W satisfies the C.P., then system (1.2) admits a G.C.L.F. at the origin. So, according to Artstein’s theorem, system (1.2) is G.S. by means of a feedback law that is smooth on $\mathbb{R}^n \setminus \{0\}$. In Theorem 3.5, we strengthen the input to state stability condition (1.8) for the subsystem $\dot{x}_1 = f_1(x_1, x_2)$ as follows: “There is a constant $c > 0$ such that, for any x with $W(x) < cV(x_1)$, we have $DV(x_1)f_1(x) < 0$, where W is a G.C.L.F. of the set M_ϕ with respect to (1.2).”

Under the previous assumption, we show that there exist a positive definite function $\tilde{V}(x_1)$ and a positive constant k so that the function

$$\Phi(x) = \tilde{V}(x_1) + kW(x)$$

is a G.C.L.F. of $O \in \mathbb{R}^n$ with respect to (1.2). Finally, we apply Theorem 3.5 to recover a result from [33].

We note that a different approach is used in [33] to investigate the global stabilizability problem for system (1.2). The sufficient conditions we propose in [33]

are also based on the existence of a G.C.L.F. of the set M_ϕ and on a suitable Lyapunov function of $O \in \mathbb{R}^{n_1}$ with respect to (1.9). The corresponding result (Theorem 3.2 in [33]) is a generalization of Artstein's theorem.

2. Almost smooth feedback stabilizers. We state a technical lemma first. The proof is similar to those given in [2] and [35], and we leave it to the reader. This lemma will be used in the proofs of the main results in this section.

LEMMA 2.1. *Suppose that there exists a continuous map $\phi: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$ and a G.C.L.F. W of the set $M_\phi = \{x \in \mathbb{R}^n: x_2 = \phi(x_1)\}$ with respect to (1.2). Then*

(a) *There exists a smooth map $r: \mathbb{R}^n \setminus M_\phi \rightarrow \mathbb{R}^m$, which satisfies the following inequality:*

$$(DWF + DWGr)(x) < -a_3(\|x_2 - \phi(x_1)\|) \quad \forall x \notin M_\phi;$$

(b) *Furthermore, the resulting map $F + Gr$ will be continuous on \mathbb{R}^n with $(Gr)(x) = 0$ for every $x \in M_\phi$, provided that W satisfies the C.P.*

THEOREM 2.2. *In addition to the hypothesis of Lemma 2.1(a), assume that*

- (i) *The set M_ϕ is positively invariant with respect to $\dot{x} = F(x)$; and*
- (ii) *System (1.10) satisfies the I.S.S.C.*

Then system (1.2) is G.S. by means of a feedback law $u = r(x)$, which is smooth on the region $\mathbb{R}^n \setminus M_\phi$.

Proof. According to Lemma 2.1(a), there exists a map $r: \mathbb{R}^n \rightarrow \mathbb{R}^m$, which is smooth for $x \notin M_\phi$, $r(x) = 0$ for $x \in M_\phi$. Furthermore, the following holds:

$$(2.1) \quad E(x) \stackrel{\text{def}}{=} -(DWF + DWGr)(x) > a_3(\|x_2 - \phi(x_1)\|) \geq a_3(\text{dist}(x, M_\phi)) > 0,$$

for every $x \notin M_\phi$. Next, we establish that the map $u = r(x)$ globally stabilizes system (1.2) at $0 \in \mathbb{R}^n$.

Since system (1.10) satisfies the I.S.S.C., there exists a Lyapunov function $V(x_1)$, which is at least continuously differentiable and uniformly unbounded on \mathbb{R}^{n_1} and which satisfies (1.8) with $f(x_1, v) \stackrel{\text{def}}{=} f_1(x_1, \phi(x_1) + v)$; namely, there exist positive constants σ and ξ such that

$$(2.2) \quad \|x_1\| > \xi, \quad \|v\| < \sigma \Rightarrow DV(x_1)f_1(x_1, \phi(x_1) + v) < 0.$$

Let c be a positive constant satisfying $a_1^{-1}(c) < \sigma$, where the function a_1 is defined in (1.5) and let

$$M_c \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n: W(x) \leq c\},$$

$$K_c \stackrel{\text{def}}{=} M_c \cap \{x \in \mathbb{R}^n: \|x_1\| > \xi\}.$$

Then it is not difficult to see that the following holds:

$$(2.3) \quad DV(x_1)f_1(x_1, x_2) < 0 \quad \forall x \in K_c.$$

Indeed, by (1.5),

$$\|x_2 - \phi(x_1)\| < a_1^{-1}(c) \quad \forall x \in K_c,$$

which, by (2.2) and the fact that $a_1^{-1}(c) < \sigma$, yields $DV(x_1)f_1(x_1, x_2) < 0$. We also define

$$M \stackrel{\text{def}}{=} \{x \in M_c: DV(x_1)f_1(x_1, x_2) \geq 0\}.$$

For each nonzero x_1 such that $\|x_1\| \leq \xi$, consider a closed sphere $S_{p_{x_1}}$ of radius $p_{x_1} > 0$ centered at $(x_1, \phi(x_1))$, which is contained in $M_c \setminus M$, and let

$$\theta_1(x_1) \stackrel{\text{def}}{=} \sup \{|DV(x_1)f_1(x_1, x_2)|, (x'_1, x'_2)' \in M_c\},$$

$$\theta_2(x_1) \stackrel{\text{def}}{=} \inf \left\{ a_3(\|x_2 - \phi(x_1)\|), (x'_1, x'_2)' \notin \bigcup_{\|x_1\| < \xi} S_{(1/2)p_{x_1}} \right\}.$$

Obviously, by (1.5), the set $M_c \cap \{x_1 \in \mathbb{R}^{n_1}: \|x_1\| \leq \xi\}$ is compact; therefore, θ_1 is upper semicontinuous and nonnegative on the region $\{x_1 \in \mathbb{R}^{n_1}: 0 < \|x_1\| \leq \xi\}$. Moreover, θ_2 is lower semicontinuous and strictly positive on this region. Therefore, there exists a real function $b: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ of class \mathcal{K} such that

$$(2.4) \quad \theta_1(x_1)b(\|x_1\|) < \theta_2(x_1) \quad \forall x_1 \neq 0: \|x_1\| \leq \xi.$$

Let ψ be any real function of class \mathcal{K}_∞ satisfying

$$(2.5) \quad V(x_1) \leq \psi(\|x_1\|) \quad \forall x_1 \in \mathbb{R}^{n_1}.$$

Finally, we define

$$(2.6) \quad \tilde{V}(x_1) = \int_0^{V(x_1)} b(\psi^{-1}(r)) dr.$$

Then, similar to [33, Thm. 3.2], we can easily establish that \tilde{V} is continuously differentiable, positive definite, and uniformly unbounded on \mathbb{R}^{n_1} . Moreover, by (2.3), we obtain

$$(2.7) \quad D\tilde{V}(x_1)f_1(x_1, x_2) = b(\psi^{-1}(V(x_1)))DV(x_1)f_1(x_1, x_2) < 0 \quad \forall x \in K_c.$$

Furthermore, it can be shown (see [33]) that the function

$$(2.8) \quad \Phi(x) = \tilde{V}(x_1) + W(x)$$

is positive definite and uniformly unbounded on \mathbb{R}^n , whereas its Lie derivative $\dot{\Phi}$ along the direction of $F + Gr$ is strictly negative on the region M_c . Indeed, for each nonzero $x \in M$ with $x_1 \neq 0$, it follows by (2.3) that $\|x_1\| < \xi$, and so, by (2.4)–(2.8), we have

$$(2.9) \quad \begin{aligned} (D\Phi(F + Gr))(x) &= b(\psi^{-1}(V(x_1)))DV(x_1)f_1(x_1, x_2) - E(x_1, x_2) \\ &\leq b(\|x_1\|)\theta_1(x_1) - \theta_2(x_1) < 0. \end{aligned}$$

For $x \in M_c \setminus M$, we have

$$DV(x_1)f(x) < 0, \quad E(x) \geq 0,$$

and so

$$(2.10) \quad (D\Phi(F + Gr))(x) = b(\psi^{-1}(V(x_1)))DV(x_1)f_1(x_1, x_2) - E(x_1, x_2) < 0.$$

Finally, from Lemma 2.1 for each nonzero $x \in M_c$ with $x_1 \neq 0$, we have

$$(2.11) \quad (D\Phi(F + Gr))(x) = -E(0, x_2) < 0.$$

From (2.9)–(2.11), it follows that there exists a function $d: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ with $d(0) = 0$ and $d(s) > 0$ for $s > 0$ such that

$$(2.12) \quad \dot{\Phi}(x) = (D\Phi(F + Gr))(x) < -d(\|x\|) \quad \forall x \in M_c \setminus \{0\}.$$

We are now in a position to prove that the map $u = v(x)$ globally stabilizes system (1.2) at $O \in \mathbb{R}^n$. First, note that the set M_c is positively invariant with respect to (1.3), where each solution $x(t, x_0)$, $x_0 \in M_c$ of (1.3) is defined (not necessarily uniquely) for all $t \geq 0$. Indeed, for each $x_0 \in M_c \setminus M_\phi$, it follows by (2.1) that $W(x(t, x_0)) \leq W(x_0) \leq c$ as long as $x(t, x_0)$ remains in $M_c \setminus M_\phi$. We distinguish two cases. The first is $x(t, x_0) \in M_c \setminus M_\phi$, and so trajectory $x(t, x_0)$ remains in M_c for every positive t , with $\|x(t, x_0)\| < +\infty$. The other case is $x(t, x_0) \in M_c \setminus M_\phi$ for any $t \in [0, T)$, where $x(T, x_0) \in M_\phi$ for some positive T . Since M_ϕ is positively invariant, for every $w \in M_\phi$, there is a trajectory $\bar{x}(t, w)$, $t \geq 0$ of the system $\dot{x} = (F + Gr)(x) = F(x)$ remaining in M_ϕ for all $t \geq 0$. Therefore, the map $\hat{x}(t, x_0) \stackrel{\text{def}}{=} \bar{x}(t - T, x(T, x_0))$ for $t \geq T$ and $\hat{x}(t, x_0) \stackrel{\text{def}}{=} x(t, x_0)$ for

$t < T$ is a solution of (1.3) starting at x_0 , which remains in M_c provided that $\|x(t, x_0)\| < +\infty$, for $t \in [0, T]$. Repeating the same discussion for any other possible solution $x(t, w)$, $w \in M_\phi$, which is defined for every positive t and does not remain in M_ϕ , we conclude that M_c is indeed a positively invariant set. Furthermore, by (1.5), (2.1), the completeness of (1.10), the continuity of ϕ , and standard Lyapunov based arguments, it can be shown that, for each $x_0 \in \mathbb{R}^n \setminus M_c$, the corresponding trajectory $x(t, x_0)$ of (1.3) enters M_c after some time $T = T(x_0)$ and remains in this region thereafter, namely,

$$(2.13) \quad x(t, x_0) \in M_c \quad \forall t \geq T.$$

(We note that, because of (1.5), (2.1) and the continuity of ϕ , the map $v(t) \doteq x_2(t, x_0) - \phi(x_1(t, x_0))$ is defined for all $t \geq 0$ with $\|x_1(t, x_0)\| < +\infty$. The latter in conjunction with the completeness of (1.10) implies that $x_1(t, x_0)$ and therefore $x(t, x_0) = (x_1'(t, x_0), x_2'(t, x_0))'$ are defined for every $t \geq 0$). Finally, by (2.12), the positively invariance of M_c , and the fact that Φ is uniformly unbounded on \mathbb{R}^n , it follows that $O \in \mathbb{R}^n$ is stable with respect to (1.3) and

$$(2.14) \quad x(t, x_0) \rightarrow 0 \quad \text{as } t \rightarrow +\infty \quad \forall x_0 \in M_c$$

for each trajectory $x(t, x_0)$ of (1.3) starting at x_0 . From (2.13) and (2.14), we conclude that $x(t, x_0) \rightarrow 0$ as $t \rightarrow +\infty$ for all $x_0 \in \mathbb{R}^n$, and so the proof is completed. \square

Example 2.3. Consider the planar case ($n = 2$) where $n_1 = n_2 = m = 1$ and

$$\begin{aligned} f_1(x_1, x_2) &= -x_1 + (x_1^2 - x_2)^2, \\ (f_2 + ug)(x_1, x_2) &= x_2 - 3x_1^2 + u(x_1^2 - x_2)^2. \end{aligned}$$

We define $\phi(x_1) = x_1^2$. Then, obviously, the system $\dot{x}_1 = f_1(x_1, \phi(x_1) + v)$ satisfies the I.S.S.C. with $V(x_1) = \frac{1}{2}x_1^2$, $\xi = 1$, and $\sigma = 1/\sqrt{2}$. Furthermore, the function $W(x_1, x_2) = \frac{1}{2}(x_2 - \phi(x_1))^2$ is a G.C.L.F. of the set $M_\phi = \{x \in \mathbb{R}^2: x_2 = \phi(x_1)\}$. In particular, $(DWG)(x) \neq 0$ for any $x = (x_1, x_2)' \notin M_\phi$. Finally, note that M_ϕ is positively invariant with respect to $\dot{x} = F(x)$. Therefore, according to Theorem 2.2, the above system is G.S. by means of an almost smooth feedback stabilizer. Finally, note that, since the linearization of the system at $O \in \mathbb{R}^2$ is completely uncontrollable and contains a strictly positive eigenvalue, stabilization via smooth feedback is impossible.

An interesting consequence of Theorem 2.2 is the following result.

THEOREM 2.4. *Suppose that there exists a G.C.L.F. W of the set*

$$M_0 \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n: x_2 = 0\}$$

with respect to (1.2), and $f_2(x_1, 0) = 0$ for all $x_1 \in \mathbb{R}^{n_1}$. Furthermore, assume that there exists a smooth map $\hat{f}_1: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$ such that

(i) *The map $\hat{f}_1(x_1)$ is homogeneous of degree $k \geq 1$, namely,*

$$\hat{f}_1(cx_1) = c^k \hat{f}_1(x_1) \text{ for all } c > 0 \text{ and } x_1 \in \mathbb{R}^{n_1};$$

(ii) *The origin $O \in \mathbb{R}^{n_1}$ is globally asymptotically stable with respect to*

$$(2.15) \quad \dot{x}_1 = \hat{f}_1(x_1);$$

(iii) *There is a positive constant ξ such that*

$$d(x_1, x_2) \stackrel{\text{def}}{=} \frac{|f_1(x_1, x_2) - \hat{f}_1(x_1)|}{\|x_1\|^k} \rightarrow 0, \quad \text{as } x_2 \rightarrow 0$$

uniformly on $x_1 \in \{x_1 \in \mathbb{R}^{n_1}: \|x_1\| > \xi\}$.

Then system (1.2) is G.S. by means of a feedback law that is smooth on $\mathbb{R}^n \setminus M_0$, provided that $\dot{x}_1 = \hat{f}_1(x_1, v)$ is complete.

Proof. Assume first that $k > 1$. Then a well-known stability result [13] asserts that conditions (i) and (ii) imply the existence of a continuously differentiable Lyapunov function $V(x_1)$ of $O \in \mathbb{R}^{n_1}$ with respect to (2.16), which is homogeneous of degree $(m-1)(k-1)$, $m > 2$. Furthermore, there are positive constants c_1 and c_2 such that the following conditions are satisfied:

$$(2.16) \quad \|DV(x_1)\| \leq c_1 \|x_1\|^{(m-1)(k-1)-1},$$

$$(2.17) \quad DV(x_1)\hat{f}_1(x_1) \leq -c_2 \|x_1\|^{m(k-1)} \quad \forall x_1 \in \mathbb{R}^{n_1}.$$

According to assumption (iii), there is a positive constant σ such that

$$(2.18) \quad d(x_1, x_2) < \frac{3}{4} \frac{c_2}{c_1}$$

for every (x_1, x_2) with $\|x_2\| < \sigma$ and $\|x_1\| > \xi$. From (2.16)-(2.18), it follows that

$$\begin{aligned} DV(x_1)f_1(x_1, x_2) &\leq |DV(x_1)f_1(x_1, x_2) - DV(x_1)\hat{f}_1(x_1)| + DV\hat{f}_1(x_1) \\ &\leq \|x_1\|^{m(k-1)}(c_1 d(x_1, x_2) - c_2) \\ &\leq -\frac{c_2}{4} \|x_1\|^{m(k-1)} < 0 \quad \forall x_1, x_2: \|x_1\| > \xi, \|x_2\| < \sigma. \end{aligned}$$

Therefore, system (1.10) satisfies the I.S.S.C. with $\phi \equiv 0$, and M_0 is positively invariant with respect to $\dot{x} = F(x)$. We conclude that the assumptions of Theorem 2.2 are satisfied, and, consequently, system (1.2) is G.S. at the origin. Finally, for $k = 1$, zero $O \in \mathbb{R}^{n_1}$ is globally exponentially stable, and so there exist a continuously differentiable real function $V(x_1)$ of $O \in \mathbb{R}^{n_1}$ with respect to (2.15) and positive constants \hat{c}_1 and \hat{c}_2 such that $\|DV(x_1)\| \leq \hat{c}_1 \|x_1\|$; $DV(x_1)\hat{f}_1(x_1) \leq -\hat{c}_2 \|x_1\|^2$ for all $x_1 \in \mathbb{R}^{n_1}$. The rest of the proof follows by using exactly the same arguments as before. \square

Example 2.5. Consider the system

$$(2.19a) \quad \dot{x}_1 = -x_1^3 + x_1^3 \theta(x_2)$$

$$(2.19b) \quad \dot{x}_2 = f_2(x_2) + u g(x_2), \quad (x_1, x_2)' \in \mathbb{R} \times \mathbb{R}^{n_2}, \quad u \in \mathbb{R},$$

where θ is continuous with $\theta(0) = 0$ and $|\theta(x_2)| \leq 1$ for all x_2 and $f_2(0) = 0$. Furthermore, assume that there exists a G.C.L.F. $W: \mathbb{R}^{n_2} \rightarrow \mathbb{R}^+$ of $O \in \mathbb{R}^{n_2}$ with respect to (2.19b). Then we can easily justify that the function $\hat{f}_1(x_1) = -x_1^3$ satisfies the assumptions of Theorem 2.4, system (2.19a) is complete, and W is a G.C.L.F. of the set M_0 with respect to (2.19). Furthermore, note that, since (2.19b) admits a G.C.L.F. at $O \in \mathbb{R}^{n_2}$, this subsystem is G.S. by means of a feedback law $u = r(x_2)$, which is smooth on $\mathbb{R}^{n_2} \setminus \{0\}$ and depends only on x_2 . This law also globally stabilizes the overall system (2.19) at $O \in \mathbb{R}^n$.

3. Existence of global control Lyapunov functions. In this section, we give sufficient conditions for the existence of global control Lyapunov functions of the origin guaranteeing stabilization by means of feedback stabilizers that are smooth on $\mathbb{R}^n \setminus \{0\}$. The following theorem generalizes Theorem 2.2 of § 2.

THEOREM 3.1. *In addition to the assumptions of Theorem 2.2, assume that the function W satisfies the C.P. Then there exists a G.C.L.F. of the origin $O \in \mathbb{R}^n$ with respect to (1.2), and so the system is G.S. by means of a feedback law $u = r(x)$, which is smooth on $\mathbb{R}^n \setminus \{0\}$.*

Proof. We proceed exactly as in the proof of Theorem 2.2. Furthermore, by Lemma 2.1(b), there exists a map $r: \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that $F + Gr$ is continuous on \mathbb{R}^n . Then,

similar to the proof of Theorem 2.2, we can establish that $O \in \mathbb{R}^n$ is globally asymptotically stable with respect to (1.3). Therefore, by Kurzweil's theorem [19], system (1.3) admits a smooth Lyapunov function $\hat{\Phi}$ at zero, which is uniformly unbounded on \mathbb{R}^n . It is not difficult to see that $\hat{\Phi}$ is a G.C.L.F. of $O \in \mathbb{R}^n$ with respect to (1.2), and so, by Artstein's theorem, the system is G.S. by means of a feedback law that is smooth on $\mathbb{R}^n \setminus \{0\}$. \square

EXAMPLE 3.2. Consider system (1.2) with $n_1 = 1$, $n_2 = 2$, $m = 1$, $x_1 = w_1 \in \mathbb{R}$, $x_2 = (w_2, w_3)' \in \mathbb{R}^2$, and

$$(3.1a) \quad f_1(x) = -w_1 - (w_2^k + w_3^l), \quad k, l \text{ integers,}$$

$$(3.1b) \quad f_2(x_2) + ug(x_2) = (-w_2, w_3 + uw_3^2)'.$$

Obviously, the system $\dot{x}_1 = f_1(x_1, 0)$ satisfies the I.S.S.C. with $V(x_1) = \frac{1}{2}x_1^2$, $\xi = 1$, and $\sigma = \frac{1}{4}$. Furthermore, the set $M_0 = \{x \in \mathbb{R}^3: x_2 = 0\}$ is positively invariant, and the function $W(x_2) = \frac{1}{2}\|x_2\|^2$ is a G.C.L.F. of M_0 with respect to the above system, which, in addition, satisfies the C.P. In particular, if we define

$$r(x_2) = \begin{cases} 0 & \text{for } w_3 = 0, \\ -2/w_3 & \text{otherwise,} \end{cases}$$

then $(DWF + rDWG)(x_2) = -w_2^2 - w_3^2$ for all $x_2 = (w_2, w_3)' \in \mathbb{R}^2$. Moreover, the resulting map rg equals $(0, -w_3)'$ for $w_3 \neq 0$ and $(0, 0)$ for $w_3 = 0$; hence it is continuous on \mathbb{R}^2 . Therefore, according to Theorem 3.1, the system is G.S. by means of a feedback law that is smooth on $\mathbb{R}^3 \setminus \{0\}$. Note that, according to our methodology, to determine the feedback stabilizer, we must first evaluate a control Lyapunov function Φ for the system $\dot{x} = F + Gr$. From a practical point of view, this is a very difficult problem. However, in our case, we can directly justify that the function

$$\Phi(x) = \frac{1}{2} w_1^2 + \frac{1}{2k} w_2^{2k} + \frac{1}{2} w_3^2$$

is a G.C.L.F. of $O \in \mathbb{R}^3$ with respect to (1.2) with dynamics given by (3.1). Therefore, we can apply Sontag's formula [24] to determine a feedback stabilizer. Indeed, Theorem 1 in [24] asserts that the feedback law

$$r(x) = \begin{cases} 0 & \text{for } x = 0, \\ -\sigma(a(x), b(x)), & \end{cases}$$

where

$$\sigma(a, b) = \begin{cases} 0 & \text{for } b = 0 \text{ and } a < 0, \\ \frac{a + \sqrt{a^2 + b^2}}{b} & \text{otherwise} \end{cases}$$

and

$$a(x) = (D\Phi F)(x) = -w_1^2 - w_1(w_2^k + w_3^l) - w_2^{2k} + w_3^2,$$

$$b(x) = (D\Phi G)(x) = w_3^3$$

is smooth on $x \in \mathbb{R}^3 \setminus \{0\}$ and globally asymptotically stabilizes the system at $O \in \mathbb{R}^3$.

An interesting consequence of Theorems 2.4 and 3.1 is the following result.

COROLLARY 3.3. *In addition to assumptions (i)–(iii) of Theorem 2.4, suppose that the mappings f_2 and g_i , $i = 1, \dots, m$, are homogeneous vector fields of degree k_0 and*

$k_i, i = 1, \dots, m$, respectively, with $k_0 > k_i$ and that they are independent of x_1 . Furthermore, there exists a positive definite matrix P_2 of dimension $n_2 \times n_2$ such that the following holds:

$$x_2' P_2 g(x_2) = 0, \quad x_2 \neq 0 \Rightarrow x_2' P_2 f_2(x_2) < 0.$$

Then there exists a G.C.L.F. of $O \in \mathbb{R}^n$ with respect to the overall system (1.2), provided that $\dot{x}_1 = f_1(x_1, v)$ is complete.

Proof. The previous assumptions in conjunction with Theorem 2 in [34] assert that the function $W(x_2) = \frac{1}{2} x_2' P_2 x_2$ is a G.C.L.F. of the set $M_0 = \{x \in \mathbb{R}^n : x_2 = 0\}$ with respect to (1.2), which, in addition, satisfies the C.P. The rest of the proof is an immediate consequence of Theorems 2.4 and 3.1. \square

Remark 3.4. Corollary 3.3 generalizes a well-known result from linear control theory concerning systems of the form

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ B \end{pmatrix} u,$$

which states that the above system is (globally) stabilizable at $O \in \mathbb{R}^n$, provided that the matrix A_{11} is Hurwitz and that the pair (A_{22}, B) is stabilizable at $O \in \mathbb{R}^{n_2}$.

Next, we establish that, if the input to state stability condition (1.8) is strengthened, it is possible to construct a G.C.L.F. of the origin with respect to (1.2). The construction of this function is based on some ideas from [32].

THEOREM 3.5. Suppose that there exist a continuous map $\phi: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$; a positive constant c ; a positive definite, continuously differentiable, real function $V: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^+$, which is uniformly unbounded on \mathbb{R}^{n_1} ; and a G.C.L.F. W of M_ϕ with respect to (1.2) satisfying the following property:

$$(3.2) \quad DV(x_1) f_1(x_1, x_2) < 0 \quad \forall x = (x_1', x_2')': W(x) < cV(x_1).$$

Then there exists a G.C.L.F. of $O \in \mathbb{R}^n$ with respect to (1.2), and so (1.2) is G.S. by means of a feedback law $u = r(x)$, which is smooth on $\mathbb{R}^n \setminus \{0\}$.

Remark 3.5. Note that the positively invariant assumption for the set M_ϕ as well as the completeness assumption for (1.10) are not required.

Proof. We establish that there exists a positive real function $s(x)$, which is smooth on $\mathbb{R}^n \setminus \{0\}$, and a positive constant k such that the function

$$(3.3) \quad \Phi(x) = s(x) V(x_1) + kW(x)$$

is a G.C.L.F. of $O \in \mathbb{R}^n$ with respect to (1.2).

We consider a smooth real function $\psi: \mathbb{R}^+ \rightarrow \mathbb{R}^+$, which is strictly decreasing on the interval $[0, 1]$ and $\psi(s) = 0$ for all $s > 1$. Furthermore, assume that its first derivative $\psi^{(1)}(s)$ is uniformly bounded on \mathbb{R}^+ , and $\psi^{(l)}(+0) = 0$ for all $l = 1, 2, \dots$. Let k be a positive real constant such that

$$(3.4) \quad k > \frac{1}{c} \sup \left\{ \left| \psi^{(1)}(s) \right|, s \in \mathbb{R}^+ \right\}$$

and define

$$s(x) \stackrel{\text{def}}{=} \begin{cases} \psi\left(\frac{1}{c} \frac{W(x)}{V(x_1)}\right) & \text{for } x_1 \neq 0, \\ 0 & \text{for } x_1 = 0. \end{cases}$$

Then the function s is smooth for $x \neq 0$ and uniformly bounded on \mathbb{R}^n . The result is that Φ is positive definite and continuous on \mathbb{R}^n and, in addition, is continuously

differentiable on $\mathbb{R}^n \setminus \{0\}$. Next, we show that Φ is uniformly unbounded on \mathbb{R}^n . Consider any sequence

$$x_n = (x'_{1n}, x'_{2n})', \quad \|x_n\| \rightarrow +\infty.$$

Without any loss of generality, we may assume that there exist subsequences $\{y_n\}$ and $\{z_n\}$ such that

$$\{x_n\} = \{y_n\} \cup \{z_n\},$$

where

$$(3.5a) \quad \|y_{2n} - \phi(y_{1n})\| \rightarrow +\infty$$

and

$$(3.5b) \quad \|z_{2n} - \phi(z_{1n})\| \leq M \quad \forall n = 1, 2, \dots$$

for some positive constant M . Then, according to (1.5), it follows that $W(y_n) \rightarrow +\infty$, and so

$$(3.6) \quad \Phi(y_n) \rightarrow +\infty \quad \text{as } \|y_n\| \rightarrow +\infty.$$

Next, consider the subsequence $\{z_n\}$. Then, since $\|z_n\| \rightarrow +\infty$ and ϕ is continuous, condition (3.5b) implies that $\|z_{1n}\| \rightarrow +\infty$. The latter, in conjunction with the fact that V is uniformly unbounded on \mathbb{R}^n , implies that $V(z_{1n}) \rightarrow +\infty$. Moreover, by (1.5) and (3.5b), we have

$$W(z_n) \leq a_2(\|z_{2n} - \phi(z_{1n})\|) \leq a_2(M) \quad \forall n = 1, 2, \dots$$

Consequently,

$$\frac{W(z_n)}{V(z_{1n})} \rightarrow 0 \quad \text{and} \quad s(z_n) = \psi\left(\frac{1}{c} \frac{W(z_n)}{V(z_{1n})}\right) \rightarrow \psi(0) > 0;$$

therefore $s(z_n)V(z_{1n}) \rightarrow +\infty$, and so

$$(3.7) \quad \Phi(z_n) = s(z_n)V(z_{1n}) + W(z_n) \rightarrow +\infty.$$

By (3.6) and (3.7), it follows that $\Phi(x_n) \rightarrow +\infty$ as $\|x_n\| \rightarrow +\infty$, which means that Φ is uniformly unbounded on \mathbb{R}^n . To complete the proof, it suffices to show that condition (1.4) is fulfilled. Indeed, consider any nonzero \hat{x} such that $(D\Phi G)(\hat{x}) = 0$. Equivalently,

$$(D\Phi G)(\hat{x}) = \left(\frac{1}{c} \psi^{(1)}\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) + k \right) (DWG)(\hat{x}) = 0,$$

and so, by (3.4), we obtain $(DWG)(\hat{x}) = 0$. Therefore, by (1.6) and (3.2), it follows that the expression

$$(3.8) \quad \begin{aligned} (D\Phi F)(\hat{x}) &= (DVf_1)(\hat{x}) \left(\psi\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) - \frac{1}{c} \psi^{(1)}\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) \frac{W(\hat{x})}{V(\hat{x}_1)} \right) \\ &\quad + (DWF)(\hat{x}) \left(\frac{1}{c} \psi^{(1)}\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) + k \right) \end{aligned}$$

is strictly negative. Indeed, for \hat{x} such that $W(\hat{x}) \geq cV(\hat{x}_1)$, we have

$$\psi\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) = \psi^{(1)}\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) = 0;$$

therefore, by (1.6) and (3.8),

$$(D\Phi F)(\hat{x}) = k(DWF)(\hat{x}) < 0.$$

Finally, for \hat{x} such that $(W(\hat{x})/V(\hat{x}_1)) < c$, it follows by (3.2) and (1.6) that $DV(\hat{x}_1)f_1(\hat{x}) < 0$ and $(DWF)(\hat{x}) \leq 0$. Furthermore, since ψ is positive and strictly decreasing on the interval $[0, 1]$, it follows that $\psi^{(1)}(s) \leq 0$ for $s \in [0, 1]$, and so

$$\psi\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) - \frac{1}{c} \psi^{(1)}\left(\frac{1}{c} \frac{W(\hat{x})}{V(\hat{x}_1)}\right) \frac{W(\hat{x})}{V(\hat{x}_1)} > 0;$$

hence by (3.8) we obtain $(D\Phi F)(\hat{x}) < 0$. Therefore, (1.4) is fulfilled, and the map Φ as defined in (3.3) is a G.C.L.F. of $O \in \mathbb{R}^n$ with respect to (1.2). Hence, according to Artstein's theorem, system (1.2) is G.S. by means of a feedback law $u = r(x)$, which is smooth on $\mathbb{R}^n \setminus \{0\}$. \square

Example 3.6. Consider the planar case

$$(3.9) \quad \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} -x_1^7 + (x_1^2 - x_2)x_1^6 \\ x_2 - x_1^2 + 2x_1^9 \end{pmatrix} + u \begin{pmatrix} 0 \\ r(x_1 + x_2) \end{pmatrix},$$

where r is a continuous function such that $r(s) = 0$ if and only if $s = 0$. We show that system (3.9) satisfies the assumptions of Theorem 3.5, and so it is G.S. at $0 \in \mathbb{R}^2$ by means of a feedback law that is smooth on $\mathbb{R}^2 \setminus \{0\}$. Let

$$f_1(x_1, x_2) = -x_1^7 + (x_1^2 - x_2)x_1^6, \quad f_2(x_1, x_2) = x_2 - x_1^2 + 2x_1^9, \quad g(x_1, x_2) = r(x_1 + x_2)$$

and define

$$\phi(x_1) = x_1^2, \quad V(x_1) = \frac{1}{2}x_1^2, \quad W(x_1, x_2) = \frac{1}{2}(x_1^2 - x_2)^2, \quad c = \frac{1}{2}.$$

Then, for any $x \in \mathbb{R}^2$ with $W(x_1, x_2) < cV(x_1)$, it holds that $|x_1^2 - x_2| < \frac{1}{2}|x_1|$, and so

$$DV(x_1)f_1(x_1, x_2) = -\frac{1}{2}x_1^8 < 0;$$

therefore (3.2) is fulfilled. Next, we show that the map W defined as above is a G.C.L.F. of the set

$$M_\phi = \{x = (x_1, x_2)' \in \mathbb{R}^2 : x_2 = x_1^2\}$$

with respect to (3.9). Indeed, for any $x \notin M_\phi$ such that $((\partial W / \partial x_2)r)(x) = 0$, we have $x_2 = -x_1$, and so

$$(DWF)(x) = -(x_1^2 + x_1)^2 < 0.$$

Therefore, according to Theorem 3.5, there exists a G.C.L.F. of $O \in \mathbb{R}^2$ with respect to (3.9).

Example 3.7. Consider the system

$$(3.10) \quad \begin{aligned} \dot{x}_1 &= -x_1^k + \theta(x), \\ \dot{x}_2 &= A_{22}x_2 + Bu, \quad x = (x_1, x_2)' \in \mathbb{R} \times \mathbb{R}^{n_2}, \end{aligned}$$

where k is odd, the function θ is continuous, and there exists a constant $M > 0$ satisfying

$$|\theta(x)| < M\|x_2\|^k \quad \forall x \in \mathbb{R} \times \mathbb{R}^{n_2}.$$

Furthermore, assume that the pair (A_{22}, B) is stabilizable. We show that system (3.10) satisfies the assumptions of Theorem 3.5; hence it admits a G.C.L.F. Indeed, since the pair (A_{22}, B) is stabilizable, there exists a positive definite matrix P_2 of dimension $n_2 \times n_2$ such that the following holds:

$$x_2'P_2B = 0 \Rightarrow x_2'P_2A_{22}x_2 < 0$$

(see [2] and [30]). Then, obviously, the function $W(x_2) = \frac{1}{2}x_2'P_2x_2$ is G.C.L.F. of the set $M_0 = \{(x_1, x_2') \in \mathbb{R} \times \mathbb{R}^{n_2}: x_2 = 0\}$ with respect to (3.10). Let p be a positive constant satisfying $x_2'Px_2 \leq p\|x_2\|^2$ for all $x_2 \in \mathbb{R}^{n_2}$. Let also $V(x_1) = \frac{1}{2}x_1^2$ and c be a positive constant with $c < 2pM^{-2/k}$. Then, for any $x = (x_1, x_2')'$ such that

$$p\|x_2\|^2 < \frac{c}{2}|x_1|^2,$$

we obtain

$$DV(x_1)(-x_1^k + \theta(x)) \leq -x_1^{k+1} + M\left(\frac{c}{2p}\right)^{2/k} x_1^{k+1} < 0,$$

and therefore (3.2) is fulfilled.

Finally, we apply Theorem 3.5 to recover an interesting result from [33].

PROPOSITION 3.8. *Suppose that there exist a continuous map $\phi: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$, a continuously differentiable Lyapunov function V of $0 \in \mathbb{R}^{n_1}$ with respect to (1.9), a G.C.L.F. $W: \mathbb{R}^n \rightarrow \mathbb{R}^+$ of M_ϕ with respect to (1.2), and positive constants c_i , $i = 1, \dots, 5$, such that*

$$(3.11) \quad V(x_1) \geq c_1\|x_1\|^2,$$

$$(3.12) \quad DV(x_1)f_1(x_1, \phi(x_1)) \leq -c_2\|x_1\|^2,$$

$$(3.13) \quad \|DV(x_1)\| \leq c_3\|x_1\|,$$

$$(3.14) \quad \|f_1(x_1, x_2) - f_1(x_1, \hat{x}_2)\| \leq c_4\|x_2 - \hat{x}_2\|,$$

$$(3.15) \quad a_2(s) \leq c_5s^2$$

for any $x_1 \in \mathbb{R}^{n_1}$, $x_2, \hat{x}_2 \in \mathbb{R}^{n_2}$, and $s \geq 0$, where a_2 is defined in (1.5). Then there exists a constant $c > 0$ such that the input to state stability condition (3.2) of Theorem 3.5 is satisfied, and so system (1.2) is G.S. by means of a feedback law that is smooth on $\mathbb{R}^n \setminus \{0\}$.

Proof. By (1.5), (3.11), and (3.15), it suffices to show that, for any $x = (x_1', x_2')' \in \mathbb{R}^n$ with

$$(3.16) \quad c_5\|x_2 - \phi(x_1)\|^2 < cc_1\|x_1\|^2,$$

it holds that

$$DV(x_1)f_1(x_1, x_2) < 0,$$

provided that

$$(3.17) \quad 0 < c \leq \frac{c_5}{c_1} \left(\frac{c_2}{c_3c_4} \right)^2.$$

Indeed, by (3.12)–(3.14), we evaluate

$$\begin{aligned} DV(x_1)f_1(x_1, x_2) &\leq |DV(x_1)f_1(x_1, x_2) - DV(x_1)f_1(x_1, \phi(x_1))| + DV(x_1)f_1(x_1, \phi(x_1)) \\ &\leq c_3c_4\|x_1\|\|x_2 - \phi(x_1)\| - c_2\|x_1\|^2. \end{aligned}$$

Therefore, by (3.16) and (3.17), we obtain

$$DV(x_1)f_1(x_1, x_2) \leq \left(c_3c_4 \sqrt{\frac{cc_1}{c_5}} - c_2 \right) \|x_1\|^2 < 0$$

for all $x \in \mathbb{R}^n$ satisfying inequality (3.16). The rest of the proof is a consequence of Theorem 3.5. \square

Example 3.9. Consider the system

$$(3.18a) \quad \dot{x}_1 = A_{11}x_1 + A_{12}x_2,$$

$$(3.18b) \quad \dot{x}_2 = f_2(x_2) + g(x_2)u, \quad (x'_1, x'_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}, \quad u \in \mathbb{R}^m,$$

where A_{11} and A_{12} are constant real matrices of dimension $n_1 \times n_1$ and $n_1 \times n_2$, respectively. Assume that

(H1) The matrix A_{11} is Hurwitz;

(H2) There is a positive definite matrix P_2 of dimension $n_2 \times n_2$ such that the function $W(x_2) = \frac{1}{2}x'_2 P_2 x_2$ is a G.C.L.F. of $O \in \mathbb{R}^{n_2}$ with respect to (3.18).

Then the assumptions of Proposition 3.8 are satisfied with $\phi = 0$, and so system (3.1) is G.S. by means of a feedback law that is smooth on $\mathbb{R}^n \setminus \{0\}$. Indeed, according to assumption (H1), there exists a positive definite matrix P_1 of dimension $n_1 \times n_1$ such that $P_1 A_{11} + A'_{11} P_1 = -I$, where I is the unit matrix of dimension $n_1 \times n_1$. Let $V(x_1) = \frac{1}{2}x'_1 P_1 x_1$. Then we can easily check that conditions (3.11)–(3.15) of Proposition 3.8 are fulfilled with V , as before, and W , as defined in (H2).

Remark 3.10. Similar to Remark 3.4, we can easily justify that Proposition 3.8 and Example 3.9 constitute generalizations of the linear case.

REFERENCES

- [1] D. AEYELS, *Stabilization of a class of nonlinear systems by a smooth feedback control*, Systems Control Lett., 5 (1985), pp. 289–294.
- [2] A. ANDREINI, A. BACCIOTTI, AND G. STEPHANI, *Global stabilizability of homogeneous vector fields of odd degree*, Systems Control Lett., 10 (1989), pp. 251–256.
- [3] Z. ARTSTEIN, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.
- [4] A. BACCIOTTI AND P. BOIERI, *Linear stabilizability of planar nonlinear systems*, Mathematics of Control Signals and Systems, 3 (1990), pp. 183–193.
- [5] S. P. BANKS, *Stabilizability of finite- and infinite-dimensional bilinear systems*, IMA J. Math. Control Inform., 1986, pp. 255–271.
- [6] W. M. BOOTHBY AND R. MARINO, *Feedback stabilization of planar nonlinear systems*, Systems Control Lett., 12 (1989), pp. 87–92.
- [7] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhauser, Boston, 1983, pp. 181–191.
- [8] C. I. BYRNES AND A. ISIDORI, *New results and counterexamples in nonlinear feedback stabilization*, Systems Control Lett., 12 (1989), pp. 437–442.
- [9] ———, *Local stabilization of minimum-phase nonlinear systems*, Systems Control Lett., 11 (1988), pp. 9–17.
- [10] P. E. CROUCH, *Spacecraft attitude control and stabilization: Applications of geometric control theory*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 321–333.
- [11] W. P. DAYAWANSA AND C. F. MARTIN, *Asymptotic stabilization of two dimensional real-analytic systems*, Systems Control Lett., 12 (1989), pp. 205–211.
- [12] P. O. GUTMAN, *Stabilizing controllers for bilinear systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 917–922.
- [13] W. HAHN, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
- [14] H. HERMES, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, SIAM J. Control Optim., 18 (1980), pp. 352–361.
- [15] V. JURDJEVIC AND J. P. QUINN, *Controllability and stability*, J. Differential Equation, 28 (1978), pp. 381–389.
- [16] N. KALOUPSIDIS AND J. TSINIAS, *Stability improvement of nonlinear systems by feedback*, IEEE Trans. Automat. Control, 29 (1984), pp. 364–367.
- [17] M. KAWSKI, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–176.
- [18] P. V. KOKOTOVIC AND H. J. SUSSMANN, *A positive real condition for global stabilization of nonlinear systems*, Systems Control Lett., 13 (1989), pp. 125–133.
- [19] J. KURZWEIL, *On the inversion of Lyapunov's second theorem on stability of motions*, Amer. Math. Soc. Transl. Ser. 2, 24 (1956), pp. 19–77.

- [20] K. K. LEE AND A. ARAPOSTATHIS, *Remarks on smooth feedback stabilization of nonlinear systems*, Systems Control Lett., 10 (1988), pp. 41–44.
- [21] R. MARINO, *Feedback stabilization of single-input nonlinear systems*, Systems Control Lett., 10 (1988), pp. 201–206.
- [22] M. SLEMROD, *Stabilization of bilinear control systems with applications to nonconservative problems in elasticity*, SIAM J. Control Optim., 16 (1978), pp. 131–141.
- [23] E. D. SONTAG, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
- [24] ———, *A “universal” construction of Artstein’s theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.
- [25] ———, *Further facts about input to state stabilization*, IEEE Trans. Automat. Control, 35 (1990), pp. 473–477.
- [26] E. D. SONTAG AND H. J. SUSSMANN, *Remarks on continuous feedback*, in Proc. IEEE Conf. Dec. and Control, Albuquerque, NM, December 1980.
- [27] ———, *Further comments on the stabilizability of the angular velocity of a rigid body*, Systems Control Lett., 12 (1989), pp. 213–217.
- [28] H. J. SUSSMANN, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.
- [29] J. TSINIAS, *Sufficient Lyapunovlike conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.
- [30] ———, *Stabilization of affine in control nonlinear systems*, Nonlinear Anal., 12 (1988), pp. 1283–1296.
- [31] ———, *Existence of control Lyapunov functions and applications to state feedback stabilizability of nonlinear systems*, SIAM J. Control Optim., 29 (1991), pp. 457–473.
- [32] ———, *Asymptotic feedback stabilization: A sufficient condition for the existence of control Lyapunov functions*, Systems Control Lett., 15 (1990), pp. 441–448.
- [33] ———, *On the existence of control Lyapunov functions: Generalizations of Vidyasagar’s theorem on nonlinear stabilization*, SIAM J. Control Optim., 30 (1992), pp. 879–893.
- [34] ———, *Remarks on feedback stabilizability of homogeneous systems*, Control Theory Adv. Tech., 6 (1990), pp. 533–541.
- [35] J. TSINIAS AND N. KALOUPSIDIS, *Output feedback stabilization*, IEEE Trans. Automat. Control, 35 (1990), pp. 951–954.
- [36] A. J. VAN DER SCHAFT, *Stabilization of Hamiltonian systems*, Nonlinear Anal., 10 (1986), pp. 1021–1035.
- [37] P. P. VARAIYA AND R. LIU, *Bounded-input bounded-output stability of nonlinear time-varying differential systems*, SIAM J. Control, 4 (1966), pp. 698–704.
- [38] M. VIDYASAGAR, *Decomposition techniques for large-scale systems with nonadditive interactions: Stability and stabilizability*, IEEE Trans. Automat. Control, 25 (1980), pp. 773–779.

KALMAN FILTERING WITH RANDOM COEFFICIENTS AND CONTRACTIONS*

PHILIPPE BOUGEROL†

Abstract. The Riccati transformation of linear filtering/control theory is shown to be a contraction on the space of positive symmetric matrices. This is used to describe the asymptotic behavior of the filter for systems with stochastic stationary parameters.

Key words. Kalman filter, stochastic parameters, Riccati equation, stationary process

AMS subject classifications. 93C05, 93E11, 60G35, 34F05

Introduction. In this paper we study the asymptotic properties of the Kalman filter in a random stationary environment, under weak controllability and observability conditions. We show that the covariance matrix of the conditional error converges in law and that the filter is exponentially stable. This is a direct generalization of Kalman's classical results.

Our main tool is the following. Consider the classical Riccati transformation that associates the error covariance matrix at time $n+1$ to the error covariance matrix at time n . We show that this transformation is a contraction with respect to the Riemannian metric on the set of positive symmetric matrices. This important fact does not seem to have been noticed before. For instance, it leads to a straightforward proof of Kalman's results on the asymptotic behavior of the filter. It can also be useful in other parts of filtering or control theory.

For the convenience of the reader who is not interested in random environment, we present our results in three parts. Section 1 is devoted to the study of the above-mentioned contraction property in the classical set-up. Filtering with random parameters is considered in §2. Section 3 is an appendix that proves the general results on iteration of random Lipschitz contractions (needed in §2).

Let us describe the main results of this paper. We consider the linear system

$$(1) \quad \begin{aligned} X_n &= A_n X_{n-1} + F_n \varepsilon_n, & n \geq 1, \\ Y_n &= C_n X_n + \eta_n, \end{aligned}$$

where $X_n \in \mathbb{R}^d$, $\varepsilon_n \in \mathbb{R}^p$, $\eta_n, Y_n \in \mathbb{R}^q$. In §1, the parameters A_n, F_n , and C_n are *deterministic* matrices with size $d \times d$, $d \times p$ and $q \times d$, respectively. The random vectors $\{(\varepsilon_n, \eta_n), n \in \mathbb{N}\}$ are independent; they have the same Gaussian law with mean 0 and covariance matrix equal to the identity. We assume that X_0 has a Gaussian law with mean \hat{X}_0 and covariance matrix P_0 . We always suppose that the matrices A_n are nonsingular (our approach does not apply in the singular case).

For any $n \geq 1$, let \mathcal{F}_n be the sigma-algebra generated by the random vectors Y_1, Y_2, \dots, Y_n , and

$$(2) \quad \hat{X}_n := \mathbb{E}(X_n / \mathcal{F}_n),$$

$$(3) \quad P_n := \mathbb{E}((X_n - \hat{X}_n)(X_n - \hat{X}_n)^* / \mathcal{F}_n).$$

* Received by the editors January 2, 1991; accepted for publication (in revised form) November 1, 1991.

† Laboratoire de Probabilités, Université Paris VI, 4, Place Jussieu, 75252 Paris Cedex 05, France.

When \mathcal{F}_n is known, \hat{X}_n is the best estimate of X_n , and P_n is the conditional error covariance matrix. Let \mathcal{P} (respectively, \mathcal{P}_0) denote the set of $d \times d$ nonnegative (respectively, positive) symmetric matrices. For any n in \mathbb{N} and $P \in \mathcal{P}$, we set

$$(4) \quad \Phi_n(P) = (A_n P A_n^* + S_n)(I + R_n S_n + R_n A_n P A_n^*)^{-1},$$

where $R_n = C_n^* C_n$ and $S_n = F_n F_n^*$; then $\Phi_n(P) \in \mathcal{P}$, and Φ_n maps \mathcal{P}_0 in \mathcal{P}_0 . The classical Kalman's recursive equations can be written as

$$(5) \quad P_n = \Phi_n(P_{n-1})$$

$$(6) \quad \hat{X}_n = (A_n - P_n R_n A_n) \hat{X}_{n-1} + P_n C_n^* Y_n$$

(see, e.g., Balakrishnan [3, Relations 4-1-19, -27, -31, -34]). The main result of § 1 is that the maps Φ_n are contractions on \mathcal{P}_0 , if we equip this set with the Riemannian metric δ , which is invariant under conjugacy (see Theorem 1.7). Moreover, the fact that these contractions are strict and/or uniform depends on the observability and the controllability properties of the linear system (1). These results are proved in their natural context, namely, by looking at the symplectic matrices that act on the set of symmetric matrices by preserving \mathcal{P}_0 . In that setting, they can be seen as generalizations of the Perron–Frobenius theorem. We also could have considered the recursion associated with $\mathbb{E}(X_n/\mathcal{F}_{n-1})$ for which analogous results hold true (see [8]).

Section 2 is devoted to filtering in a random stationary environment. We consider again the linear equation (1), but we now suppose that the parameters A_n , F_n , and C_n are *stochastic* and that $\{(A_n, F_n, C_n), n \geq 1\}$ is a stationary ergodic process. Under suitable hypotheses (see Hypothesis in § 2), system (1) is conditionally Gaussian, and \hat{X}_n and P_n are also given by the recursive equations (5) and (6) of Kalman. These hypotheses hold, for instance, when the parameters are independent of the noises. We first describe in § 2.1 some actual situations that can be described by such systems. Then, in § 2.2 we introduce weak controllability and observability assumptions (in contrast with the uniform conditions of Kalman). These conditions can hold for systems that are usually neither controllable nor observable. Fault-tolerant systems usually have this property. Under these assumptions, we describe the asymptotic behavior of the conditional error covariance matrices P_n . Our main result is Theorem 2.4. We show that there exists a stationary \mathcal{P}_0 -valued process $\{\bar{P}_n, n \in \mathbb{N}\}$ with the following universal property: “Almost surely, for any solution P_n of (5), $\|P_n - \bar{P}_n\|$ converges to 0 as $n \rightarrow +\infty$.” In particular, P_n converges in law. In § 2.3 we prove that the filter (6) is exponentially stable. These results are deduced from properties of processes that are defined by iterations under stationary Lipschitz maps (Relation (5) and Theorem 1.7 show that the process P_n is of this type). These properties are interesting for their own sake; they are established in § 3.

This paper is self-contained and in some sense elementary. Some of its ideas are already in the literature. The trick of studying the filtering of Riccati's equation through the action of symplectic matrices is, of course, well known (see, e.g., Hermann [15], Shayman [24], and their references). Our semigroup \mathcal{H} has been introduced already by Wojtkowski [31], [32] in a different context. The contraction property of the Riccati transformation is a generalization of the contraction property of matrices with nonnegative elements for the Hilbert metric, due to G. Birkhoff [5]. It's also related with the contraction properties of product of random matrices on boundaries. In Bougerol [7] we recover some of the results obtained here by making use of the Osseledets theorem and the Lyapunov exponents of the associated Hamiltonian matrices.

All our results can be generalized easily to the continuous-time case, either by a direct study or by a reduction to discrete-time.

1. Contraction properties of Riccati equation.

1.1. The semigroup of Hamiltonian matrices. We consider the classical linear system (1) with deterministic parameters (A_n, F_n, C_n) , $n \geq 1$,

$$X_n = A_n X_{n-1} + F_n \varepsilon_n,$$

$$Y_n = C_n X_n + \eta_n,$$

defined in the Introduction. We always suppose that the matrices A_n are invertible. We associate to this system the so-called Hamiltonian matrices M_n of order $2d$ written in block form as

$$(7) \quad M_n = \begin{pmatrix} A_n & S_n A_n^{*-1} \\ R_n A_n & (I + R_n S_n) A_n^{*-1} \end{pmatrix},$$

where $R_n = C_n^* C_n$ and $S_n = F_n F_n^*$. These matrices are in the symplectic group $\text{Sp}(d, \mathbb{R})$. This group is defined as the set of all the matrices M of order $2d$ such that $M^* J M = J$, where $J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$, (I is the identity matrix of order d and M^* is the transpose of M). This relation can be written as $M^{-1} = J M^* J$, thus we see that M^* is also in $\text{Sp}(d, \mathbb{R})$. If we write

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where the entries are $d \times d$ matrices, then BA^* and A^*C are symmetric and $A^*D - C^*B = I$.

Let \mathcal{P} (respectively, \mathcal{P}_0) be the set of $d \times d$ nonnegative (respectively, positive) symmetric matrices (we recall that a matrix M is nonnegative, respectively, positive, when for all $x \neq 0$, $x^* M x \geq 0$, respectively, > 0). The set of all Hamiltonian matrices is

$$\mathcal{H} = \left\{ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \text{Sp}(d, \mathbb{R}); A \text{ is invertible, } BA^* \in \mathcal{P}, A^*C \in \mathcal{P} \right\}.$$

Indeed, every matrix M_n defined previously is in \mathcal{H} , and every matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ in \mathcal{H} is the Hamiltonian matrix of the linear system (1) with the constant parameters $A_n = A$, $F_n = \sqrt{BA^*}$, $C_n = \sqrt{CA^{-1}}$, and dimensions $p = q = d$. We define three subsets \mathcal{H}_1 , \mathcal{H}_2 , and \mathcal{H}_0 of \mathcal{H} by

$$\mathcal{H}_1 = \left\{ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathcal{H}; BA^* \in \mathcal{P}, A^*C \in \mathcal{P}_0 \right\},$$

$$\mathcal{H}_2 = \left\{ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathcal{H}; BA^* \in \mathcal{P}_0, A^*C \in \mathcal{P} \right\},$$

$$\mathcal{H}_0 = \mathcal{H}_1 \cap \mathcal{H}_2.$$

We remark that \mathcal{H}_2 is the dual of \mathcal{H}_1 (in the sense that $M \in \mathcal{H}_1$ if and only if $M^* \in \mathcal{H}_2$). The following semigroup property of \mathcal{H} already appeared in a different context, implicitly in Ol'shanskii [22] and explicitly in Wojtkowski [31] and [32].

PROPOSITION 1.1. *The product of matrices in \mathcal{H} is in \mathcal{H} . The product of a matrix in \mathcal{H} with a matrix in \mathcal{H}_1 (respectively, \mathcal{H}_2 , \mathcal{H}_0) is in \mathcal{H}_1 (respectively, \mathcal{H}_2 , \mathcal{H}_0). In other words, \mathcal{H} is a semigroup of matrices, and \mathcal{H}_1 , \mathcal{H}_2 , and \mathcal{H}_0 are two-sided ideals of \mathcal{H} .*

We will need the following well-known lemma.

LEMMA 1.2. *When P and Q are in \mathcal{P} , then $I + PQ$ is invertible.*

Proof. If P is positive, we can find an invertible matrix M such that $P^{-1} = M^*M$ and $Q = M^*DM$, where D is diagonal with nonnegative entries. Then $I + PQ = M^{-1}(I + D)M$. Thus, the eigenvalues of $I + PQ$ are greater than one. By density, this remains true even if P is only nonnegative. \square

Proof of Proposition 1.1. Let M_1 and M_2 be two matrices in \mathcal{H} , and $M_3 = M_2M_1$. For $i = 1, 2$, or 3 we write $M_i = \begin{pmatrix} A_i & B_i \\ C_i & D_i \end{pmatrix}$, and we set $Q_1 = C_1A_1^{-1}$ and $P_2 = A_2^{-1}B_2$. Since $M_1 \in \mathcal{H}$, the matrix $A_1^*C_1$ is in \mathcal{P} and the fact that $Q_1 = A_1^{*-1}(A_1^*C_1)A_1^{-1}$ yields that Q_1 is in \mathcal{P} . Similarly, P_2 is also in \mathcal{P} . We have

$$A_3 = A_2A_1 + B_2C_1 = A_2(I + A_2^{-1}B_2C_1A_1^{-1})A_1 = A_2(I + P_2Q_1)A_1.$$

Hence, it follows from the previous lemma that A_3 is invertible. We will make use of the relation

$$A_2^*D_2 = C_2^*B_2 + I = C_2^*A_2P_2 + I = A_2^*C_2P_2 + I$$

in the next computation. Since $C_3 = C_2A_1 + D_2C_1$, one has

$$\begin{aligned} A_3^*C_3 &= A_1^*(I + Q_1P_2)A_2^*(C_2A_1 + D_2C_1) \\ &= A_1^*(I + Q_1P_2)(A_2^*C_2A_1 + A_2^*C_2P_2C_1 + C_1) \\ (8) \quad &= A_1^*(I + Q_1P_2)A_2^*C_2(A_1 + P_2C_1) + A_1^*C_1 + C_1^*P_2C_1 \\ &= A_1^*(I + Q_1P_2)A_2^*C_2(I + P_2Q_1)A_1 + A_1^*C_1 + C_1^*P_2C_1. \end{aligned}$$

This shows that $A_3^*C_3$ is a nonnegative symmetric matrix. Similarly (or using transpositions) we see that $A_3B_3^*$ is also nonnegative. This proves that M_3 is in \mathcal{H} . Thus, \mathcal{H} is a semigroup. If, moreover, $A_2^*C_2$ or $A_1^*C_1$ is invertible, then (8) shows that $A_3^*C_3$ is positive definite. Hence, \mathcal{H}_2 is an ideal of \mathcal{H} . Similarly, \mathcal{H}_1 , and thus \mathcal{H}_0 , is also an ideal of \mathcal{H} .

The following result will be useful later to link the fact that a linear system is observable or controllable with the fact that the associated Hamiltonian matrices are in \mathcal{H}_1 or in \mathcal{H}_2 .

PROPOSITION 1.3. *Let $M_n = \begin{pmatrix} A_n & B_n \\ C_n & D_n \end{pmatrix}$, $n \in \mathbb{N}$, be matrices in \mathcal{H} . Then $M_nM_{n-1} \dots M_1$ is in \mathcal{H}_1 if and only if*

$$\text{Det}(A_1^*C_1 + A_1^*A_2^*C_2A_1 + \dots + A_1^* \dots A_{n-1}^*A_n^*C_nA_{n-1} \dots A_1) \neq 0,$$

and $M_nM_{n-1} \dots M_1$ is in \mathcal{H}_2 if and only if

$$\text{Det}(B_nA_n^* + A_nB_{n-1}^*A_{n-1}^*A_n^* + \dots + A_n \dots A_2B_1A_1^*A_2^* \dots A_n^*) \neq 0.$$

Proof. For any $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ in \mathcal{H} , let $\xi(M) = A^*C$ and $\alpha(M) = A$. We first show that, if M_1 and M_2 are in \mathcal{H} , then for any x in \mathbb{R}^d ,

$$(9) \quad \xi(M_2M_1)x = 0$$

if and only if

$$(10) \quad \xi(M_1)x = 0 \quad \text{and} \quad \xi(M_2)A_1x = 0.$$

We use the following two properties:

- (i) When $P, Q \in \mathcal{P}$ and $(P + Q)x = 0$, then $Px = Qx = 0$;
- (ii) For any matrix M , if $P \in \mathcal{P}$ and $M^*PMx = 0$, then $PMx = 0$.

We know by (8) that if $Q_1 = C_1A_1^{-1}$ and $P_2 = A_2^{-1}B_2$, then

$$(11) \quad \xi(M_2M_1) = A_1^*(I + Q_1P_2)A_2^*C_2(I + P_2Q_1)A_1 + A_1^*C_1 + C_1^*P_2C_1.$$

The right-hand side is a sum of nonnegative symmetric matrices. Therefore, we see that if (9) holds, then $A_1^* C_1 x = 0$ and $A_2^* C_2 (I + P_2 Q_1) A_1 x = 0$. This implies that $C_1 x = 0$ (since A_1 is invertible) and $A_2^* C_2 A_1 x = 0$ (since $Q_1 A_1 x = C_1 x = 0$); thus, (10) holds. The converse also follows easily from (11). Using the equivalence between (9) and (10) we see that the following statements are equivalent:

$$\begin{aligned} & \xi(M_n \dots M_2 M_1) x = 0, \\ \Leftrightarrow & \xi(M_1) x = 0 \quad \text{and} \quad \xi(M_n \dots M_2) \alpha(M_1) x = 0, \\ \Leftrightarrow & \dots, \\ \Leftrightarrow & \xi(M_1) x = 0 \quad \text{and} \quad \xi(M_2) \alpha(M_1) x = 0, \dots, \\ \text{and} & \quad \xi(M_n) \alpha(M_{n-1}) \dots \alpha(M_1) x = 0. \end{aligned}$$

Since the matrices $\xi(M_n)$ are nonnegative and $\alpha(M_n)$ invertible, this is also equivalent to

$$\begin{aligned} & \{ \xi(M_1) + \alpha(M_1)^* \xi(M_2) \alpha(M_1) + \dots + \alpha(M_1)^* \dots \alpha(M_{n-1})^* \xi(M_n) \\ & \cdot \alpha(M_{n-1}) \dots \alpha(M_1) \} x = 0. \end{aligned}$$

This proves the first claim. The second claim is obtained by duality.

1.2. Contraction property. For any matrix $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ in \mathcal{H} we define a map $\Phi_M : \mathcal{P}_0 \rightarrow \mathcal{P}_0$ by

$$(12) \quad \Phi_M(T) = (AT + B)(CT + D)^{-1}, \quad T \in \mathcal{P}_0.$$

The fact that the right-hand side is a well-defined element of \mathcal{P}_0 will be shown in Proposition 1.5. A straightforward computation shows that the map $M \rightarrow \Phi_M$ defines an action of the semigroup \mathcal{H} on \mathcal{P}_0 , in the sense that, for any M, N in \mathcal{H} ,

$$(13) \quad \Phi_{MN} = \Phi_M \circ \Phi_N$$

(in fact this action is induced by the linear action of symplectic matrices on d -dimensional linear subspaces of \mathbb{R}^{2d}). We remark that the relation (4), which defines the maps Φ_n , can be written as

$$\Phi_n(P) = (A_n P + S_n A_n^{*-1})(R_n A_n P + (I + R_n S_n) A_n^{*-1})^{-1}.$$

This shows that $\Phi_n = \Phi_{M_n}$. Therefore, by (5), the error covariance matrix P_n satisfies

$$(14) \quad P_n = \Phi_{M_n}(P_{n-1}).$$

This equation is called the *discrete Riccati equation*. The classical continuous matrix Riccati equation on \mathcal{P}_0 is similar. Indeed, under some mild regularity and boundedness assumptions, if P_t , $t \in \mathbb{R}^+$, is the solution of the matrix-valued differential equation

$$\dot{P}_t = A_t P_t + P_t A_t^* - P_t R_t P_t + S_t, \quad P_0 \in \mathcal{P}_0,$$

where R_t , S_t , $t \geq 0$, are in \mathcal{P} , then there exists a family N_t , $t \geq 0$, of matrices in \mathcal{H} such that $P_t = \Phi_{N_t}(P_0)$ (cf., Hermann [15]).

DEFINITION 1.4. The Riemannian distance δ on \mathcal{P}_0 is defined by: for any $P, Q \in \mathcal{P}_0$,

$$\delta(P, Q) = \left\{ \sum_{i=1}^d \text{Log}^2 \lambda_i \right\}^{1/2},$$

where $\lambda_1, \dots, \lambda_d$ are the eigenvalues of the matrix PQ^{-1} .

It is shown in Maass [19, Thm. p. 27] (see also Terras [27]), that δ is the usual Riemannian distance on \mathcal{P}_0 when this set is considered as the Riemannian symmetric

space $\text{Gl}(d, \mathbb{R})/O(d)$ (this metric is associated to the arc length $ds^2 = \text{tr} \{(P^{-1} dP)^2\}$, which is invariant under conjugacy and coincides with the Euclidean arc length on the logarithms of the diagonal matrices in \mathcal{P}_0). In particular, (\mathcal{P}_0, δ) is complete and δ induces the usual topology. The main property of this distance is its invariance under conjugacy and under inversion. For any invertible matrix A and for all P, Q in \mathcal{P}_0 ,

$$\delta(APA^*, AQA^*) = \delta(P, Q) = \delta(P^{-1}, Q^{-1}).$$

We next prove that the transformations Φ_M are contractions of (\mathcal{P}_0, δ) when $M \in \mathcal{H}$.

PROPOSITION 1.5. *Let $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ be a matrix in \mathcal{H} . Then*

(i) *For any T in \mathcal{P} (respectively, \mathcal{P}_0), $CT + D$ is invertible and $(AT + B)(CT + D)^{-1}$ is in \mathcal{P} (respectively, \mathcal{P}_0).*

(ii) *If $M \in \mathcal{H}_2$, then for any T in \mathcal{P} , $(AT + B)(CT + D)^{-1}$ is in \mathcal{P}_0 .*

Proof. Let $T \in \mathcal{P}$. The matrices $P = A^{-1}B$, $Q = CA^{-1}$, and $S = A(T + P)A^*$ are in \mathcal{P} . Since

$$CT + D = QAT + QAP + A^{*-1} = (QS + I)A^{*-1},$$

it follows from Lemma 1.2 that $CT + D$ is invertible. Now, the relation

$$(AT + B)(CT + D)^{-1} = (AT + AP)A^*(QS + I)^{-1} = S(QS + I)^{-1}$$

easily implies the proposition (we note that this is equal to $(S^{-1} + Q)^{-1}$ when S is invertible). \square

We always use the Euclidean norm on \mathbb{R}^d and the associated operator norm on the set of matrices: if M is a matrix of order d , we let $\|M\| = \text{Sup}\{\|Mx\|; x \in \mathbb{R}^d, \|x\| = 1\}$.

PROPOSITION 1.6. *Let T, S be matrices in \mathcal{P}_0 and $\alpha = \text{Max}(\|T\|, \|S\|)$. Then for all $P \in \mathcal{P}$,*

$$\delta(T + P, S + P) \leq \frac{\alpha}{\alpha + \beta} \delta(T, S)$$

where $\beta = \text{Inf}\{\langle Px, x \rangle; \|x\| = 1\}$.

Proof. The mean value theorem yields that, when $0 < a, b \leq m$ and $r > 0$, then

$$(15) \quad \text{Log} \frac{a+r}{b+r} \leq \frac{m}{m+r} \text{Log}^+ \frac{a}{b}$$

(where $\text{Log}^+ x = \text{Max}(\text{Log} x, 0)$). It is known and not difficult to prove (see Gantmacher [14, Ch. 10, § 7]) that the eigenvalues of TS^{-1} are real, positive, and that they have the following Min-Max representation. Let

$$\lambda_1(T, S) \leq \lambda_2(T, S) \leq \dots \leq \lambda_d(T, S)$$

be the eigenvalues of TS^{-1} written in ascending order. Then

$$\lambda_k(T, S) = \text{Min} \left\{ \text{Max} \left\{ \frac{\langle Tx, x \rangle}{\langle Sx, x \rangle}; x \in V \right\}; V \in \Gamma(k) \right\},$$

where $\Gamma(k)$ is the set of k -dimensional linear subspaces of \mathbb{R}^d . We prove that

$$(16) \quad |\text{Log} \lambda_k(T + P, S + P)| \leq \frac{\alpha}{\alpha + \beta} |\text{Log} \lambda_k(T, S)|,$$

for any $1 \leq k \leq d$. We first suppose that $\lambda_k(T + P, S + P) > 1$. Relation (15) entails that

$$\text{Log} \frac{\langle Tx, x \rangle + \langle Px, x \rangle}{\langle Sx, x \rangle + \langle Px, x \rangle} \leq \frac{\alpha}{\alpha + \beta} \text{Log}^+ \frac{\langle Tx, x \rangle}{\langle Sx, x \rangle}.$$

Thus,

$$\begin{aligned}
 |\operatorname{Log} \lambda_k(T+P, S+P)| &= \operatorname{Log} \lambda_k(T+P, S+P) \\
 &= \operatorname{Log} \operatorname{Min} \left\{ \operatorname{Max} \left\{ \frac{\langle (T+P)x, x \rangle}{\langle (S+P)x, x \rangle}; x \in V \right\}; V \in \Gamma(k) \right\} \\
 &= \operatorname{Min} \left\{ \operatorname{Max} \left\{ \operatorname{Log} \frac{\langle Tx, x \rangle + \langle Px, x \rangle}{\langle Sx, x \rangle + \langle Px, x \rangle}; x \in V \right\}; V \in \Gamma(k) \right\} \\
 &\leq \frac{\alpha}{\alpha + \beta} \operatorname{Min} \left\{ \operatorname{Max} \left\{ \operatorname{Log}^+ \frac{\langle Tx, x \rangle}{\langle Sx, x \rangle}; x \in V \right\}; V \in \Gamma(k) \right\} \\
 &\leq \frac{\alpha}{\alpha + \beta} \operatorname{Log}^+ \lambda_k(T, S).
 \end{aligned}$$

Since the left-hand term is positive, this first gives that $\operatorname{Log}^+ \lambda_k(T, S) = |\operatorname{Log} \lambda_k(T, S)|$, and then that (16) holds. When $\lambda_k(T+P, S+P) = 1$, (16) is obvious. When $\lambda_k(T+P, S+P) < 1$, we use the relation $\lambda_k(T+P, S+P) = 1/\lambda_{d-k+1}(S+P, T+P)$, and we apply (16) to $\lambda_{d-k+1}(S+P, T+P)$. Finally, (16) implies immediately the proposition. \square

THEOREM 1.7. *The following properties hold:*

(i) *For any M in \mathcal{H} , and T, S in \mathcal{P}_0 ,*

$$\delta(\Phi_M(T), \Phi_M(S)) \leq \delta(T, S).$$

(ii) *For any M in \mathcal{H}_1 or in \mathcal{H}_2 , and T, S in \mathcal{P}_0 ,*

$$\delta(\Phi_M(T), \Phi_M(S)) < \delta(T, S).$$

(iii) *For any M in \mathcal{H}_0 , there exists $\rho(M)$, $0 < \rho(M) < 1$, such that, for all T, S in \mathcal{P}_0 ,*

$$\delta(\Phi_M(T), \Phi_M(S)) \leq \rho(M) \delta(T, S).$$

Proof. Let $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ be a matrix in \mathcal{H} . The matrices $P = A^{-1}B$ and $Q = CA^{-1}$ are in \mathcal{P} and

$$M = \begin{pmatrix} I & 0 \\ Q & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & A^{*-1} \end{pmatrix} \begin{pmatrix} I & P \\ 0 & I \end{pmatrix}.$$

We consider the transformations $\tau_P(T) = T + P$, $\tau_Q(T) = T + Q$, $\gamma_A(T) = ATA^*$ and $\sigma(T) = T^{-1}$ defined on \mathcal{P}_0 . By making use of (13) we obtain

$$(17) \quad \Phi_M(T) = (\sigma \circ \tau_Q \circ \sigma \circ \gamma_A \circ \tau_P)(T).$$

We have already noticed that γ_A and σ are isometries of the metric space (\mathcal{P}_0, δ) . It follows from Proposition 1.6 that τ_P and τ_Q are contractions. This and (17) prove (i). If M is in \mathcal{H}_1 , then Q is invertible. Hence, τ_Q is a strict contraction by Proposition 1.6. Similarly, when M is in \mathcal{H}_2 , P is invertible and τ_P is a strict contraction. Thus, (ii) follows from (17). Let us prove (iii). We consider an M in \mathcal{H}_0 . Then both P and Q are invertible. Moreover, for any T in \mathcal{P}_0 , $\tau_P(T) \geq P$ (in the sense that $\tau_P(T) - P \in \mathcal{P}$), which implies that $(\gamma_A \circ \tau_P)(T) \geq APA^*$, and thus,

$$(\sigma \circ \gamma_A \circ \tau_P)(T) \leq (APA^*)^{-1}.$$

Let $\zeta = \|(APA^*)^{-1}\|$ and $\varepsilon = \inf \{\langle Qx, x \rangle; \|x\| = 1\}$. It follows from Proposition 1.6 that

for all T_1 and T_2 in \mathcal{P}_0 ,

$$\begin{aligned} & \delta(\tau_Q[(\sigma \circ \gamma_A \circ \tau_P)(T_1)], \tau_Q[(\sigma \circ \gamma_A \circ \tau_P)(T_2)]) \\ & \leq \frac{\zeta}{\zeta + \varepsilon} \delta((\sigma \circ \gamma_A \circ \tau_P)(T_1), (\sigma \circ \gamma_A \circ \tau_P)(T_2)) \\ & \leq \frac{\zeta}{\zeta + \varepsilon} \delta(T_1, T_2). \end{aligned}$$

Since σ is an isometry, this relation and (16) yield that (iii) holds with $\rho(M) = \zeta/(\zeta + \varepsilon)$. \square

As an application, let us outline a short proof of a classical result of Kalman. We suppose that the linear system with constant coefficients

$$X_n = AX_{n-1} + F\varepsilon_n, \quad Y_n = CX_n + \eta_n,$$

is controllable and observable. Let M be the Hamiltonian matrix (here independent of n) defined by (7). It follows from Proposition 1.3 that there is an integer $p > 0$ such that M^p is in \mathcal{H}_0 . Therefore, by Theorem 1.7, Φ_M has a power that is a uniform contraction. This implies by the fixed point theorem that there exists a matrix P in \mathcal{P}_0 such that all the solutions P_n of (5) converge to P when $n \rightarrow +\infty$, as soon as P_0 is in \mathcal{P}_0 .

2. Filtering with random parameters. We now study the asymptotic behavior of linear filtering in a random environment. We consider the case where the parameters A_n , F_n , C_n of the linear equation (1) are stochastic. More precisely, we suppose that the following hypothesis holds.

Hypothesis H. For all $n \geq 1$, the quantities A_n , F_n , C_n , ε_n , η_n are random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\{(A_n, F_n, C_n), n \geq 1\}$ is a strictly stationary ergodic process. There is a σ -algebra \mathcal{F}_0 contained in \mathcal{F} such that, if $\mathcal{F}_n = \sigma(\mathcal{F}_0, Y_1, \dots, Y_n)$ is the σ -algebra generated by \mathcal{F}_0 and by Y_1, Y_2, \dots, Y_n , then for all $n \geq 1$,

- (i) A_n , F_n , and C_n are \mathcal{F}_{n-1} measurable.
- (ii) The random vector (ε_n, η_n) has a Gaussian law with mean zero and covariance matrix equal to the identity matrix. It is independent of X_{n-1} and \mathcal{F}_{n-1} .
- (iii) Conditionally on \mathcal{F}_0 , the random vector X_0 has a Gaussian law with mean \hat{X}_0 and covariance matrix P_0 .

This set-up is called *conditionally Gaussian*. The conditional expectations $\hat{X}_n = \mathbb{E}(X_n / \mathcal{F}_n)$ and the conditional error covariance matrices $P_n = \mathbb{E}((X_n - \hat{X}_n)(X_n - \hat{X}_n)^* / \mathcal{F}_n)$ are given by the recursions (5) and (6) (see, e.g., Whittle [29, p. 260]). Work on such systems with stochastic parameters goes back to Kalman [17], [18]. A recent reference is De Koning [12] (see also Nahi [21]). An important example is the following: suppose that $\{(\varepsilon_n, \eta_n), n \geq 1\}$ is a sequence of independent normalized Gaussian random vectors, independent of a stationary ergodic process $(A_n, F_n, C_n), n \geq 1$. Then, if $\mathcal{F}_0 = \sigma\{(A_n, F_n, C_n), n \geq 1\}$, the hypothesis (H) holds.

In § 2.1, we present some examples of real situations that can be modeled by these equations. In § 2.2, we describe the asymptotic behavior of P_n as $n \rightarrow +\infty$. The exponential stability of the filter is proved in § 2.3. We will always suppose that the matrices A_n are nonsingular. Without loss of generality, we can and will suppose that the stationary process (A_n, F_n, C_n) is defined on $(\Omega, \mathcal{F}, \mathbb{P})$ for all $n \in \mathbb{Z}$.

2.1. Examples.

2.1.1. Filter with periodic parameters. We suppose that there exist functions A , B , C on $\Omega_1 = \mathbb{Z}/p\mathbb{Z}$ such that, for all $\omega \in \Omega_1$,

$$(A_n(\omega), B_n(\omega), C_n(\omega)) = (A(\omega + n), B(\omega + n), C(\omega + n)),$$

where $\omega + n$ is the sum modulo p . Let \mathcal{F}_1 be the set of all the subsets of Ω_1 , and let \mathbb{P}_1 be the uniform measure on Ω_1 . We also consider a sequence of noises (ε_n, η_n) , $n \geq 1$, defined on some $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$. Then these coefficients define a linear system on the probability space $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$, $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$, for which (H) holds. These systems with periodic parameters have been studied recently, for instance, by De Souza and Goodwin [13], and Bittanti, Colaneri, and Di Nicolao [6].

2.1.2. Random sampling. In several situations, a linear system can be observed only at random times $T_0 < T_1 < \dots$. This so-called stochastic sampling phenomenon can occur because of technical imperfections in the instrumentation. It may also be applied intentionally, for instance, when a digital computer is time shared in a stochastic manner as suggested by Kalman [17]. In Snyder and Fishman [25], the tracking of fireflies, which can be observed only by their flashes, is studied (we can easily imagine some more realistic examples); see also Chang [11]. These systems are used in modeling of ARMA processes with missing data.

The basic model is the usual time-invariant system

$$(18) \quad X_n = AX_{n-1} + F\varepsilon_n, \quad Y_n = CX_n + \eta_n.$$

We suppose that the state is observed only at random times T_n , $n \geq 0$, independent of this system, and that $\{T_{n+1} - T_n, n \geq 0\}$ is a stationary ergodic process with values in \mathbb{N}^* . If we let $Z_n = X_{T_n}$ and $W_n = Y_{T_n}$ then

$$Z_n = A^{T_n - T_{n-1}} Z_{n-1} + \sum_{k=0}^{T_n - T_{n-1} - 1} A^k F \varepsilon_{T_n - k}, \quad W_n = CZ_n + \eta_{T_n}.$$

For each $n \geq 1$, let $A_n = A^{T_n - T_{n-1}}$ and let F_n be a symmetric matrix such that

$$F_n^2 = \sum_{k=0}^{T_n - T_{n-1} - 1} A^k F F^* A^{*k}.$$

Using if necessary a generalized inverse of F_n , it is easy to see that there exists a sequence of independent Gaussian random variables $\alpha_n \in \mathbb{R}^p$, $\beta_n \in \mathbb{R}^q$, with mean 0 and covariance matrix equal to the identity, independent of the sequence $\{(A_n, F_n), n \geq 1\}$, such that

$$\sum_{k=0}^{T_n - T_{n-1} - 1} A^k F \varepsilon_{T_n - k} = F_n \alpha_n, \quad \eta_{T_n} = \beta_n.$$

We obtain that

$$(19) \quad Z_n = A_n Z_{n-1} + F_n \alpha_n, \quad W_n = CZ_n + \beta_n.$$

This is a system with stochastic parameters for which (H) holds. In this setting, the asymptotic properties of the filter have been studied by Viano [28] under the additional assumption that the matrix A is stable and that C^*C is invertible. We are able to treat the case where the system (18) is only controllable and observable. We remark that no uniform controllability property of (19) can be expected when the T_n 's are not bounded. When the random variables $T_{n+1} - T_n$ are independent and identically distributed, the error covariance P_n , $n \in \mathbb{N}$, is a Markov chain on \mathcal{P}_0 . But this process is singular (it does not satisfy the Harris irreducibility condition). Even in that case, it does not seem easy to study its asymptotic behavior without recourse to contractions properties.

2.1.3. Fault-tolerant filtering. Consider a failure-prone linear system. It can be, for instance, a manufacturing plant or a space-station under the bombardment of

meteorites. We can assume that the plant state has two equations: at time n , either $X_n = MX_{n-1} + \varepsilon_n$ if the system is operational, or $X_n = NX_{n-1} + \varepsilon_n$ if the system is in a state of failure and undergoing repair. The failure/repair process may be modeled by a stationary sequence A_n , $n \geq 1$, of random matrices such that $A_n \in \{M, N\}$. Such systems are considered, for instance, in Akella and Kumar [1] and in Mariton [20] (see also Willems and Willems [30]). If $Y_n = CX_n + \eta_n$, the associated filtering system will satisfy (H).

We may also consider a filtering system with a failure-prone observation process. This can be due to the instrumentation or to the fact that at some unexpected times, the state cannot be determined. For instance, we can think of the tracking of a plane, which is sometimes hidden by clouds. A model for this situation can be

$$X_n = AX_{n-1} + F\varepsilon_n, \quad Y_n = C_nX_n + \eta_n,$$

where C_n is equal to some matrix C when the observation process is operational and some other matrix D , otherwise. Notice that it is natural to assume that under failure, i.e., when $C_n = D$, the system is not observable.

It follows from the results of the next section that the filter has very good asymptotic properties. This shows that in some sense, Kalman's filtering is fault tolerant. Of course, users are already aware of this fact.

2.1.4. Estimation of AR processes with AR parameters. Suppose that we observe an univariate autoregressive (AR) process $Z_n = \rho_n Z_{n-1} + \eta_n$, where the parameters ρ_n satisfy $\rho_n = a\rho_{n-1} + \varepsilon_n$. Here, $\{(\varepsilon_n, \eta_n), n \geq 1\}$ is a sequence of independent normalized Gaussian random variables. These models occur, for instance, in stochastic adaptive control (see, e.g., Caines and Meyn [10]). If we want to estimate the parameter a , it is useful to compute the conditional law of ρ_n , once Z_1, \dots, Z_n are observed. Let $X_n = \rho_n$, $Y_n = Z_n$, $A_n = a$, $F_n = 1$, and $C_n = Z_{n-1}$. This system can be written as (1). If $a \in (-1, 1)$, then the parameter sequence is stationary and Hypothesis (H) holds.

2.2. Asymptotic properties of the error covariance matrix. We consider a linear system (1) with random parameters for which (H) holds. In particular, the process (A_n, F_n, C_n) is stationary and ergodic. In the sequel, δ is the distance on \mathcal{P}_0 introduced in Definition 1.4. We recall that $R_n = C_n^* C_n$ and $S_n = F_n^* F_n$. For any $n \geq 1$, let

$$\Omega_n = \{\omega \in \Omega; \text{Det}(A_1^* R_1 A_1 + A_1^* A_2^* R_2 A_2 A_1 + \dots + A_1^* \dots A_n^* R_n A_n \dots A_1) \neq 0\},$$

and

$$\Xi_n = \{\omega \in \Omega; \text{Det}(S_n + A_n S_{n-1} A_n^* + \dots + A_n \dots A_2 S_1 A_2^* \dots A_n^*) \neq 0\}.$$

DEFINITION 2.1. The system (1) is called *weakly observable* if for some $n > 0$, $\mathbb{P}(\Omega_n) > 0$; it is called *weakly controllable* if for some $n > 0$, $\mathbb{P}(\Xi_n) > 0$.

When the parameters are deterministic, we recover the usual observability and controllability conditions. But these notions are much weaker than the one commonly used in the study of time-dependent systems (see, e.g., Jazwinski [16], Anderson and Moore [2]). In some of the examples given in § 2.1, only these weak conditions were natural. We will need the following lemmas.

LEMMA 2.2. Let M_n , $n \in \mathbb{N}$, be the sequence of Hamiltonian matrices associated with a linear system (1) satisfying (H). If the system is weakly observable (respectively, weakly controllable), then, almost surely, $M_n \dots M_1$ is in \mathcal{H}_1 (respectively, \mathcal{H}_2) for all $n \in \mathbb{N}$ large enough.

Proof. If the system is weakly observable, then $\mathbb{P}(\Omega_k) > 0$ for some $k \geq 1$. Proposition 1.3 yields that $M_k(\omega)M_{k-1}(\omega) \dots M_1(\omega)$ is in \mathcal{H}_1 when $\omega \in \Omega_k$. Thus, $\mathbb{P}(M_k \dots M_1 \in \mathcal{H}_1) > 0$. It follows from the ergodic theorem that for almost all $\omega \in \Omega$ there exists an integer p , depending on ω , such that $M_{p+k}(\omega) \dots M_{p+1}(\omega) \in \mathcal{H}_1$. Since \mathcal{H}_1 is an ideal in \mathcal{H} (cf. Proposition 1.1) this shows that, almost surely, for n large enough, $M_n \dots M_1 \in \mathcal{H}_1$. When the system is weakly controllable, the proof is similar. \square

LEMMA 2.3. *For any $Q \in \mathcal{P}_0$,*

$$(20) \quad \text{Max} (\text{Log}^2 \|Q\|, \text{Log}^2 \|Q^{-1}\|) \leq \delta(Q, I)^2 \leq d \text{Max} (\text{Log}^2 \|Q\|, \text{Log}^2 \|Q^{-1}\|).$$

Proof. If $\lambda_1 \leq \dots \leq \lambda_d$ are the eigenvalues of Q , then $\lambda_1 = 1/\|Q^{-1}\|$ and $\lambda_d = \|Q\|$. Since $\delta(Q, I)^2 = \sum_{i=1}^d \text{Log}^2 \lambda_i$, the conclusion of the lemma is clear. \square

Our main result is the following theorem. It implies that:

(i) The filtering process is successful, since the conditional error covariance matrix P_n does not explode. This error is asymptotically stationary. For instance, it converges in law (see Corollary 3.3).

(ii) Even for a fixed $\omega \in \Omega$ (outside an exceptional subset of measure 0), there is no optimal choice of the initial condition P_0 , since all the sequences P_n have the same asymptotic behavior. An analogous result for the usual distance on \mathcal{P} is shown in Proposition 2.5.

THEOREM 2.4. *We consider a linear system (1) with stochastic parameters for which*

(a) *Condition (H) holds.*

(b) *The system is weakly observable and weakly controllable.*

(c) *The random variables $\text{LogLog}^+ \|A_1\|$, $\text{LogLog}^+ \|A_1^{-1}\|$, $\text{LogLog}^+ \|C_1\|$, $\text{LogLog}^+ \|F_1\|$ are integrable.*

Then, there exists an ergodic stationary \mathcal{P}_0 -valued process $\{\bar{P}_n, n \in \mathbb{Z}\}$, that is solution of (5). Furthermore, there is a negative real number $\alpha < 0$ such that, almost surely, for any solution P_n of (5) for which $P_0 \in \mathcal{P}_0$,

$$\overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log} \delta(P_n, \bar{P}_n) \leq \alpha < 0.$$

Proof. We are going to apply Theorem 3.1, proved in the Appendix, to the sequence $\{\Phi_n, n \in \mathbb{Z}\}$ of random contractions of the metric space (\mathcal{P}_0, δ) defined by (4). We first check the condition (C1) of this theorem, namely that for some P in \mathcal{P}_0 , $\mathbb{E}[\text{Log} \delta(\Phi_1(P), P)]$ is finite. Actually, we choose P equal to the identity matrix I . Let $T = A_1 A_1^* + S_1$. We get $\Phi_1(I) = T(I + R_1 T)^{-1} = (T^{-1} + R_1)^{-1}$. Since $T - (T^{-1} + R_1)^{-1}$ is a nonnegative matrix, we have

$$\|\Phi_1(I)\| \leq \|(T^{-1} + R_1)^{-1}\| \leq \|T\| \leq \|A_1 A_1^*\| + \|S_1\| \leq \|A_1\|^2 + \|F_1\|^2$$

and

$$\|\Phi_1(I)^{-1}\| \leq \|T^{-1} + R_1\| \leq \|T^{-1}\| + \|R_1\| \leq \|(A_1 A_1^*)^{-1}\| + \|R_1\| \leq \|A_1^{-1}\|^2 + \|C_1\|^2.$$

By Lemma 2.3,

$$(21) \quad \delta(\Phi_1(I), I)^2 \leq d \text{Max} (\text{Log}^2 \|\Phi_1(I)\|, \text{Log}^2 \|\Phi_1(I)^{-1}\|).$$

Using these inequalities and hypothesis (c) from Theorem 2.4, we see that $\mathbb{E}[\text{Log} \delta(\Phi_1(I), I)]$ is finite. Now we check that Condition (C2) holds. By Theorem 1.7, Φ_n is a contraction. Thus, it suffices to show that the coefficient of contraction $\rho(\Phi_p \circ \dots \circ \Phi_1)$ is smaller than 1 for some $p > 0$, with positive probability. Let M_n , $n \in \mathbb{N}$, be the Hamiltonian matrices associated to (1). It follows from Lemma 2.2 that, almost surely, for all $p \in \mathbb{N}$ large enough, $M_p \dots M_1$ is both in \mathcal{H}_1 and in \mathcal{H}_2 . Since

$\mathcal{H}_0 = \mathcal{H}_1 \cap \mathcal{H}_2$, this yields that for some p , $\mathbb{P}(M_p \dots M_1 \in \mathcal{H}_0) \neq 0$. By (12), $\Phi_p \circ \dots \circ \Phi_1 = \Phi_{M_p \dots M_1}$, and therefore, $\mathbb{P}(\rho(\Phi_p \circ \dots \circ \Phi_1) < 1) \neq 0$ by Theorem 1.7 (iii). Thus, Condition (C2) of Theorem 3.1 holds. This theorem implies the result. \square

PROPOSITION 2.5. *We suppose that the hypotheses of Theorem 2.4 hold and that $\text{Log}^+ \|A_1\|$, $\text{Log}^+ \|A_1^{-1}\|$, $\text{Log}^+ \|C_1\|$, and $\text{Log}^+ \|F_1\|$ are integrable. Then, almost surely, for any solution P_n of (5) for which $P_0 \in \mathcal{P}_0$,*

$$(22) \quad \overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log} \|P_n - \bar{P}_n\| \leq \alpha < 0.$$

Proof. Since P_n and \bar{P}_n are symmetric positive matrices, we can find matrices K_n and D_n such that $\bar{P}_n = K_n^* K_n$, $P_n = K_n^* D_n K_n$, and such that D_n is a diagonal matrix with positive entries $\lambda_1^{(n)}, \dots, \lambda_d^{(n)}$. We have $\delta(P_n, \bar{P}_n) = \{\sum_{i=1}^d \text{Log}^2 \lambda_i^{(n)}\}^{1/2}$. It follows from Theorem 2.4 that

$$\overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log} |\text{Log} \lambda_i^{(n)}| \leq \alpha,$$

for $i = 1, \dots, d$. This implies that

$$(23) \quad \overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log} |1 - \lambda_i^{(n)}| \leq \alpha.$$

As in the proof above, we see that $\mathbb{E}[\delta(\Phi_1(I), I)]$ is finite by (21). Moreover, by Lemma 2.3, $\text{Log} \|\bar{P}_n\| \leq \delta(I, \bar{P}_n)$. Thus, it follows from Proposition 3.4 that, almost surely,

$$(24) \quad \overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log} \|\bar{P}_n\| \leq 0.$$

For each $x \in \mathbb{R}^d$, $\|K_n x\|^2 = \langle K_n^* K_n x, x \rangle \leq \|\bar{P}_n x\|$, so that $\|K_n\|^2 \leq \|\bar{P}_n\|$. This yields that

$$\|P_n - \bar{P}_n\| \leq \|K_n^* (D_n - I) K_n\| \leq \|K_n\|^2 \|D_n - I\| \leq \|\bar{P}_n\|^2 \max_{1 \leq i \leq d} |1 - \lambda_i^{(n)}|.$$

It is clear that (22) is a consequence of (23), (24), and of this inequality. \square

Remark. It is not difficult to see that the conclusion of the theorem also holds when P_0 is only in \mathcal{P} . (The main point is to note that since the system is weakly controllable, there is almost surely an integer k such that $M_k \dots M_1$ is in \mathcal{H}_2 , by Lemma 2.2, which implies that $P_k = \Phi_{M_k \dots M_1}(P_0)$ is in \mathcal{P}_0 , by Proposition 1.5 (ii).)

2.3. Stability of the filter. We show that the linear equation (6) of the filter is exponentially stable. We make use of the classical method of Lyapunov, as in Anderson and Moore [2], for instance.

THEOREM 2.6. *We consider a system (1) with stochastic parameters for which:*

- (i) *Hypothesis (H) holds.*
- (ii) *The system is weakly observable and weakly controllable.*
- (iii) *The random variables $\text{Log}^+ \|A_1\|$, $\text{Log}^+ \|A_1^{-1}\|$, $\text{Log}^+ \|C_1\|$, $\text{Log}^+ \|F_1\|$ are integrable. Then the equation (6) of the filter is exponentially stable: namely, there is a real number $\gamma > 0$ such that, almost surely,*

$$\overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log} \|(A_n - P_n R_n A_n) \dots (A_1 - P_1 R_1 A_1)\| \leq -\gamma < 0,$$

for any solution $\{P_n, n \in \mathbb{N}\}$ of (5) such that $P_0 \in \mathcal{P}_0$.

We need the following classical lemma. It is an immediate consequence of the relation

$$\hat{X}_n - X_n = (A_n - P_n R_n A_n)(\hat{X}_{n-1} - X_{n-1}) + P_n C_n^* \eta_n + (P_n R_n - I) F_n \varepsilon_n,$$

which itself results from (1) and (6).

LEMMA 2.7. Let $B_n = A_n - P_n R_n A_n$ and $T_n = P_n R_n P_n + (I - P_n R_n) S_n (I - P_n R_n)^*$. Then, $P_n = B_n P_{n-1} B_n^* + T_n$.

LEMMA 2.8. Let $G_n = T_n + B_n T_{n-1} B_n^* + \dots + B_n \dots B_2 T_1 B_2^* \dots B_n^*$. Under the hypotheses of Theorem 2.6, there exists $n \in \mathbb{N}$ such that $\mathbb{P}(\text{Det}(G_n) \neq 0) > 0$.

Proof. We follow an argument in Anderson and Moore [2], where a similar result is proved. For each integer $i \geq 1$, let $K_i = P_i C_i^*$ and $H_i = (I - P_i R_i)$. We consider the two $d \times (p+q)n$ matrices

$$W_n = (K_n, F_n, A_n K_{n-1}, A_n F_{n-1}, \dots, A_n A_{n-1} \dots A_2 K_1, A_n A_{n-1} \dots A_2 F_1)$$

$$V_n = (K_n, H_n F_n, B_n K_{n-1}, B_n H_{n-1} F_{n-1}, \dots, B_n B_{n-1} \dots B_2 K_1, B_n B_{n-1} \dots B_2 H_1 F_1).$$

It is easy to see by straightforward manipulations that there exists a $(p+q)n \times (p+q)n$ upper triangular matrix U_n , with all diagonal terms equal to 1, such that $V_n = W_n U_n$. We remark that

$$W_n W_n^* \geq S_n + A_n S_{n-1} A_n^* + \dots + A_n \dots A_2 S_1 A_2^* \dots A_n^*.$$

Since the system is weakly controllable, there exists a positive integer n such that the subset Ξ_n , where the right-hand side is invertible, is of positive measure. On Ξ_n , the rank of W_n is d . The same property holds for V_n since $V_n = W_n U_n$ and since U_n is invertible. Then the lemma results from the fact that $G_n = V_n V_n^*$. \square

Proof of Theorem 2.6. For notational simplicity we suppose that $\mathbb{P}(\text{Det}(G_1) \neq 0) > 0$. The general case is treated in the same way (by looking at the sequence P_{nk} , $k \in \mathbb{N}$, where n is given by Lemma 2.8). For any $n \in \mathbb{N}$, let $\lambda_n = \|T_n^{-1}\|^{-1}$, $\sigma_n = \|P_n\|$, $\alpha = \|P_0^{-1}\|$, with the convention that $\lambda_n = 0$, if T_n is not invertible. Let p be a positive integer. For a fixed $x_p \in \mathbb{R}^d$, we define a finite sequence x_0, x_1, \dots, x_p , by the backward recursion $x_n = B_{n+1}^* x_{n+1}$. Let $V_n = x_n^* P_n x_n$. We have

$$V_{n+1} - V_n = x_{n+1}^* T_{n+1} x_{n+1} \geq \lambda_{n+1} \|x_{n+1}\|^2 \geq \frac{\lambda_{n+1}}{\sigma_{n+1}} V_{n+1}$$

so that, if $\tau_n = (1 - \lambda_n / \sigma_n)$, then $V_n \leq \tau_{n+1} V_{n+1}$. Therefore,

$$\|x_0\|^2 \leq \|P_0^{-1}\| V_0 \leq \|P_0^{-1}\| \tau_1 \dots \tau_p V_p \leq \|P_0^{-1}\| \tau_1 \dots \tau_p \|P_p\| \|x_p\|^2.$$

Since $x_0 = B_1^* \dots B_p^* x_p$, this implies that

$$\|B_p \dots B_1\|^2 \leq \alpha \tau_1 \dots \tau_p \sigma_p$$

and

$$\frac{1}{p} \text{Log} \|B_p \dots B_1\|^2 \leq \frac{1}{p} \text{Log} \alpha + \frac{1}{p} \text{Log} \sigma_p + \frac{1}{p} \sum_{i=1}^p \text{Log} \tau_i.$$

As in the proof of (24) we can apply Proposition 3.4 to see that, almost surely,

$$\overline{\lim}_{p \rightarrow +\infty} \frac{1}{p} \text{Log} \sigma_p \leq 0.$$

Therefore, it follows from Birkhoff's ergodic theorem (see the proof of Corollary 3.2) that, almost surely,

$$\overline{\lim}_{p \rightarrow +\infty} \frac{1}{p} \text{Log} \|B_p \dots B_1\|^2 \leq \mathbb{E}(\text{Log } \tau_1).$$

Since we have supposed that $\mathbb{P}(\text{Det}(G_1) \neq 0) > 0$, we know that $\mathbb{E}(\text{Log } \tau_1) < 0$. \square

Remark 1. The exponential rate γ can be chosen to be equal to the smallest positive Lyapunov exponent of the associated Hamiltonian matrices. This is shown in Bougerol [7]. 2. It is not difficult to see that the theorem also holds if we only suppose that $P_0 \in \mathcal{P}$.

3. Appendix. Iteration of stationary Lipschitz functions. In this appendix, we establish some general properties of the processes that are obtained by iteration of random Lipschitz functions. At least in particular cases, similar results are already known. But we think that our set-up and formulation can be useful in several situations; we applied them in § 2.

Let (E, δ) be a complete separable metric space. A Lipschitz map $\phi: E \rightarrow E$ is a map for which

$$\rho(\phi) := \text{Sup} \left\{ \frac{\delta(\phi(x), \phi(y))}{\delta(x, y)}; x, y \in E, x \neq y \right\}$$

is finite. If ϕ and φ are such Lipschitz maps, then

$$(25) \quad \rho(\phi \circ \varphi) \leq \rho(\phi)\rho(\varphi).$$

When $\rho(\phi) \leq 1$, the map ϕ is called a *contraction*. It is called a *uniform contraction* when $\rho(\phi) < 1$. We consider a stationary ergodic process $\{\phi_n, n \in \mathbb{Z}\}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where each $\phi_n: E \rightarrow E$ is a random Lipschitz map (we suppose that the maps $(\omega, x) \in \Omega \times E \rightarrow \phi_n''(x) \in E$ are measurable when E is equipped with its Borel σ -algebra; for notational convenience, we do not write ω explicitly). We consider the processes $X_n, n \in \mathbb{N}$, on E for which the following difference equation holds:

$$(26) \quad X_n = \phi_n(X_{n-1}).$$

The following theorem is more or less known. It generalizes results of Sunyach [26], Brandt [9], and Barnsley and Elton [4]. We recall that $\text{Log}^+ x = \text{Max}(\text{Log } x, 0)$. If E' is a countable dense subset of E , then $\rho(\phi_1)$ is the supremum of the countable set $\{\delta(\phi_1(x), \phi_1(y))/\delta(x, y); x, y \in E', x \neq y\}$; thus, $\rho(\phi_1)$ is measurable.

THEOREM 3.1. *Let $\{\phi_n, n \in \mathbb{Z}\}$ be a stationary ergodic sequence of Lipschitz maps from E into E . We suppose that the following conditions hold:*

(C1) *For some x in E , $\mathbb{E}[\text{Log}^+ \delta(\phi_1(x), x)]$ is finite.*

(C2) *The random variable $\text{Log}^+ \rho(\phi_1)$ is integrable, and for some integer $p > 0$, the real number*

$$\alpha = \frac{1}{p} \mathbb{E}[\text{Log } \rho(\phi_p \circ \phi_{p-1} \circ \dots \circ \phi_1)]$$

is strictly negative.

Then there exists an ergodic stationary process $\{\tilde{X}_n, n \in \mathbb{Z}\}$ with values in E , solution of (26), such that, almost surely,

$$\overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log } \delta(X_n, \tilde{X}_n) \leq \alpha < 0$$

for any process $\{X_n, n \geq 0\}$, such that $X_n = \phi_n(X_{n-1})$ for all $n > 0$.

Proof. Let us first show that, almost surely,

$$(27) \quad \overline{\lim}_{k \rightarrow +\infty} \frac{1}{k} \operatorname{Log} \rho(\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k}) \leq \alpha.$$

By (25),

$$\overline{\lim}_{k \rightarrow +\infty} \frac{1}{k} \operatorname{Log} \rho(\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k}) \leq \overline{\lim}_{k \rightarrow +\infty} \frac{1}{k} \operatorname{Log} \rho(\phi_{-1} \circ \cdots \circ \phi_{-k}),$$

thus, the left-hand side of (27) is a subinvariant function. By ergodicity, it is constant, almost surely. Let β be this constant. We know that for any nonnegative integrable random variable Z , $\sum_{k=0}^{+\infty} \mathbb{P}(Z > k)$ is finite. Therefore, the integrability condition in (C2) entails that, for any $\varepsilon > 0$, $\sum_{k=0}^{+\infty} \mathbb{P}(\operatorname{Log}^+ \rho(\phi_1) > k\varepsilon) < +\infty$. Since all the ϕ_n 's have the same law, this entails that $\sum_{k=0}^{+\infty} \mathbb{P}(\operatorname{Log}^+ \rho(\phi_{-k}) > k\varepsilon) < +\infty$, so that by the Borel-Cantelli lemma, almost surely,

$$\overline{\lim}_{k \rightarrow +\infty} \frac{1}{k} \operatorname{Log} \rho(\phi_{-k}) \leq 0.$$

By making use of this inequality, and of the fact that if $k = mp + r$, where r is an integer in $[0, p)$,

$$\begin{aligned} \operatorname{Log} \rho(\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k}) &\leq \left\{ \sum_{i=0}^{m-1} \operatorname{Log} \rho(\phi_{-ip} \circ \phi_{-ip-1} \circ \cdots \circ \phi_{-(i+1)p+1}) \right\} \\ &\quad + \operatorname{Log} \rho(\phi_{-mp} \circ \cdots \circ \phi_{-k}), \end{aligned}$$

we see that

$$\beta \leq \overline{\lim}_{k \rightarrow +\infty} \frac{1}{mp} \sum_{i=0}^{m-1} \operatorname{Log} \rho(\phi_{-ip} \circ \phi_{-ip-1} \circ \cdots \circ \phi_{-(i+1)p+1}).$$

It follows from Birkhoff's ergodic theorem that the expectation of the right-hand side is equal to α , proving (27). On the other hand, it follows as above from the integrability condition (C1) and from the Borel-Cantelli lemma that for some fixed x in E , almost surely,

$$(28) \quad \overline{\lim}_{k \rightarrow +\infty} \frac{1}{k} \operatorname{Log} \delta(\phi_{-k}(x), x) \leq 0.$$

Now

$$\begin{aligned} &\delta((\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k})(x), (\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k} \circ \phi_{-k-1})(x)) \\ &\leq \rho(\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k}) \delta(x, \phi_{-k-1}(x)), \end{aligned}$$

thus, (27) and (28) imply that, almost surely,

$$\overline{\lim}_{k \rightarrow +\infty} \frac{1}{k} \operatorname{Log} \delta((\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k})(x), (\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k} \circ \phi_{-k-1})(x)) \leq \alpha.$$

Since $\alpha < 0$, this shows that $\{(\phi_0 \circ \phi_{-1} \circ \cdots \circ \phi_{-k})(x), k \in \mathbb{N}\}$ is a Cauchy sequence for almost all $\omega \in \Omega$. We suppose that E is complete, thus, this sequence converges. In the same way, we see that $(\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_{n-k})(x)$ converges, almost surely, for each fixed $n \in \mathbb{Z}$, when $k \rightarrow +\infty$. Let

$$(29) \quad \tilde{X}_n = \lim_{k \rightarrow +\infty} (\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_{n-k})(x).$$

Since $\{\phi_k, k \in \mathbb{Z}\}$ is a stationary ergodic process, and since we can write $\tilde{X}_n = F(\phi_k, k \leq n)$ for some measurable function F independent of $n \in \mathbb{Z}$, we see that \tilde{X}_n is itself a stationary ergodic process. It is clear that it is a solution of (26). Finally, let $\{X_n, n \geq 0\}$ be any process satisfying (26). Since

$$\begin{aligned} \delta(X_n, \tilde{X}_n) &= \delta((\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1)(X_0), (\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1)(\tilde{X}_0)) \\ &\leq \rho(\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1) \delta(X_0, \tilde{X}_0), \end{aligned}$$

we see that, almost surely,

$$\overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \text{Log } \delta(X_n, \tilde{X}_n) \leq \frac{1}{p} \mathbb{E}(\text{Log } \rho(\phi_p \circ \phi_{p-1} \circ \cdots \circ \phi_1)) = \alpha.$$

This concludes the proof of the theorem. \square

Remark. Suppose that $\text{Log}^+ \rho(\phi_1)$ is integrable, then if $\mathbb{E}\{\text{Log } \delta(\phi_1(x_0), x_0)\}$ is finite for some x_0 in E , then it is finite for all $x \in E$ since

$$\begin{aligned} \delta(\phi_1(x), x) &\leq \delta(\phi_1(x), \phi_1(x_0)) + \delta(\phi_1(x_0), x_0) + \delta(x_0, x) \\ &\leq \delta(\phi_1(x_0), x_0) + (\rho(\phi_1) + 1) \delta(x_0, x). \end{aligned}$$

COROLLARY 3.2. *Let π be the common law of the \tilde{X}_n 's. Under the hypotheses above, almost surely, for any sequence X_n satisfying (26), and for any bounded continuous function $f: E \rightarrow \mathbb{R}$,*

$$\lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n f(X_i) = \int f d\pi.$$

Proof. The hypotheses on E imply that there exists a countable set D of bounded continuous functions on E such that any sequence of probability measures μ_n on E converges weakly to a probability measure μ if and only if $\int f d\mu_n \rightarrow \int f d\mu$ for any f in D (see, e.g., Parthasarathy [23, Thm. II.6.6]). Let $\Omega(f)$ be the subset of Ω , where $1/n \sum_{i=1}^n f(\tilde{X}_i)$ converges to $\int f d\pi$. Since (\tilde{X}_n) is stationary and ergodic, $\mathbb{P}(\Omega(f)) = 1$ by Birkhoff's ergodic theorem. Let Ω_0 be the intersection of the set, where $\delta(X_n, \tilde{X}_n)$ converges to 0 and of all the sets $\Omega(f)$, $f \in D$. It follows from Theorem 3.1 that $\mathbb{P}(\Omega_0) = 1$. We fix an ω in Ω_0 . Let m_n be the empirical measure of the sequence $\{\tilde{X}_n(\omega), n \geq 1\}$, defined by $\int f dm_n = 1/n \sum_{i=1}^n f(\tilde{X}_i(\omega))$, when $f: E \rightarrow \mathbb{R}$ is bounded and continuous. The sequence $\{m_n, n \geq 1\}$ converges weakly to π . Moreover, since $\delta(X_n(\omega), \tilde{X}_n(\omega)) \rightarrow 0$,

$$(30) \quad \lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n f(X_i(\omega)) = \lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^n f(\tilde{X}_i(\omega)) = \int f d\pi$$

when f is uniformly continuous. This, in turn, yields that the empirical measure of the sequence $\{X_n(\omega), n \geq 1\}$ converges weakly to π , i.e., that (30) holds for any function f that is bounded and continuous (see Parthasarathy [23, Thm. II.6.1]). \square

COROLLARY 3.3. *Under the hypotheses of the previous theorem, all the solutions X_n of (26) converge in law to the same limit π .*

This corollary is an immediate consequence of the theorem. The following technical proposition has been used in § 2.

PROPOSITION 3.4. *Suppose that the random maps ϕ_n are contractions, that $\mathbb{E}[\delta(\phi_1(x), x)]$ is finite for some x in E , and that for some $p > 0$, $\mathbb{E}[\text{Log } (\rho(\phi_p \circ \cdots \circ \phi_1))] < 0$. Then, almost surely,*

$$\lim_{n \rightarrow +\infty} \frac{1}{n} \delta(x, X_n) = 0$$

for any sequence $\{X_n, n \geq 0\}$ for which (26) holds.

Proof. For any $n \geq 1$,

$$\begin{aligned} \delta(x, (\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1)(x)) &\leq \delta(x, (\phi_n \circ \cdots \circ \phi_2)(x)) \\ &\quad + \delta((\phi_n \circ \cdots \circ \phi_2)(x), (\phi_n \circ \cdots \circ \phi_1)(x)) \\ &\leq \delta(x, (\phi_n \circ \cdots \circ \phi_2)(x)) + \rho(\phi_n \circ \cdots \circ \phi_2) \delta(x, \phi_1(x)). \end{aligned}$$

So that, by induction,

$$(31) \quad \delta(x, (\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1)(x)) \leq \sum_{i=1}^n \rho(\phi_n \circ \cdots \circ \phi_{i+1}) \delta(x, \phi_i(x)).$$

Since the ϕ_n 's are contractions, this implies in particular, using (25), that

$$\delta(x, (\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1)(x)) \leq \delta(x, \phi_n(x)) + \sum_{i=1}^{n-1} \rho(\phi_{i+1}) \delta(x, \phi_i(x)).$$

Thus, by Birkhoff's ergodic theorem, almost surely,

$$\overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \delta(x, (\phi_n \circ \cdots \circ \phi_1)(x)) \leq \overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^{n-1} \rho(\phi_{i+1}) \delta(x, \phi_i(x)) \leq \mathbb{E}[\rho(\phi_1) \delta(x, \phi_0(x))].$$

Similarly, for any fixed k in \mathbb{N} , we obtain from (31) that, when $n > k$,

$$\begin{aligned} \delta(x, (\phi_n \circ \cdots \circ \phi_1)(x)) &\leq \sum_{i=n-k+1}^n \rho(\phi_n \circ \cdots \circ \phi_{i+1}) \delta(x, \phi_i(x)) \\ &\quad + \sum_{i=1}^{n-k} \rho(\phi_{i+k} \circ \cdots \circ \phi_{i+1}) \delta(x, \phi_i(x)) \end{aligned}$$

and by the ergodic theorem,

$$(32) \quad \begin{aligned} \overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \delta(x, (\phi_n \circ \cdots \circ \phi_1)(x)) &\leq \overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \sum_{i=1}^{n-1} \rho(\phi_{i+k} \circ \cdots \circ \phi_{i+1}) \delta(x, \phi_i(x)) \\ &\leq \mathbb{E}[\rho(\phi_k \circ \cdots \circ \phi_1) \delta(x, \phi_0(x))]. \end{aligned}$$

Now, $\rho(\phi_k \circ \cdots \circ \phi_1) \delta(x, \phi_0(x))$ converges, almost surely, to 0 as $k \rightarrow +\infty$ and is dominated by the integrable function $\delta(x, \phi_0(x))$. Thus, its expectation goes to 0 by Lebesgue's theorem; from (32) we obtain

$$\overline{\lim}_{n \rightarrow +\infty} \frac{1}{n} \delta(x, (\phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1)(x)) = 0$$

almost surely. Finally, since $X_n = (\phi_n \circ \cdots \circ \phi_1)(X_0)$, we see that

$$\begin{aligned} \delta(x, X_n) &\leq \delta(x, (\phi_n \circ \cdots \circ \phi_1)(x)) + \delta((\phi_n \circ \cdots \circ \phi_1)(x), (\phi_n \circ \cdots \circ \phi_1)(X_0)) \\ &\leq \delta(x, (\phi_n \circ \cdots \circ \phi_1)(x)) + \delta(x, X_0), \end{aligned}$$

so that

$$\delta(x, X_n)/n \rightarrow 0 \quad \text{as} \quad n \rightarrow +\infty. \quad \square$$

Remark. When the random maps ϕ_n are independent and identically distributed, then $\mathbb{E}(\delta(\tilde{X}_n, x))$ is finite under the hypothesis of this proposition. In that case, its conclusion follows directly from the Borel-Cantelli lemma.

REFERENCES

- [1] R. AKELLA AND P. R. KUMAR, *Optimal control production rate in a failure prone manufacturing system*, IEEE Trans. Automat. Control, AC-31 (1985), pp. 116-126.
- [2] B. D. O. ANDERSON AND J. B. MOORE, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20-32.
- [3] A. V. BALAKRISHNAN, *Kalman filtering theory*, Optimization Software, Inc., New York, 1984.
- [4] M. F. BARNESLEY AND J. H. ELTON, *A new class of Markov processes for image encoding*, Adv. Appl. Probab., 20 (1988), pp. 14-32.
- [5] G. BIRKHOFF, *Extensions of Jentzsch's theorem*, Trans. Amer. Math. Soc., 85 (1957), pp. 219-227.
- [6] S. BITTANTI, P. COLANERI, AND G. DE NICOLAO, *The difference periodic Riccati equation for the periodic prediction problem*, IEEE Trans Automat. Control, AC-33 (1988), pp. 706-712.

- [7] PH. BOUGEROL, *Filtre de Kalman Bucy et exposants de Lyapounov*, in Lyapunov Exponents, Proceedings, Oberwolfach 1990, L. Arnold, H. Crauel, and J. P. Eckmann, eds., Lecture Notes in Math., Springer-Verlag, Berlin, Heidelberg, New York, 1486 (1991), pp. 112–122.
- [8] ———, *Some results on the filtering equation with random parameters*, in Applied Stochastic Analysis, I. Karatzas and D. Ocone, eds., Lecture Notes in Control, Springer-Verlag, Berlin, Heidelberg, New York, 177 (1992), pp. 30–37.
- [9] A. BRANDT, *The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients*, Adv. Appl. Probab., 18 (1986), pp. 211–220.
- [10] P. E. CAINES AND S. P. MEYN, *A new approach to stochastic adaptative control*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 220–226.
- [11] S. S. CHANG, *Optimum filtering and control of randomly sampled systems*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 537–546.
- [12] W. L. DE KONING, *Optimal estimation of linear discrete-time systems with stochastic parameters*, Automatica, 20 (1984), pp. 113–115.
- [13] C. E. DE SOUZA AND G. C. GOODWIN, *Periodic solutions of matrix Riccati equation in optimal filtering of nonstabilizable periodic systems*, Proc. 10th World IFAC Congress, R. Ishermann, ed., Pergamon Press, IX (1987), pp. 243–248.
- [14] F. R. GANTMACHER, *The theory of matrices*, Vol. 1, Chelsea Press, New York, 1959.
- [15] R. HERMANN, *Cartanian geometry, nonlinear waves, and control theory*, Part A, Interdisciplinary Mathematics, Vol. 20, Math Sci. Press, Brookline, MA, 1979.
- [16] A. H. JAZWINSKI, *Stochastic processes and filtering theory*, Academic Press, New York, 1970.
- [17] R. E. KALMAN, *Analysis and synthesis of linear systems operating on randomly sampled data*, Ph.D. thesis, Columbia University, New York, 1957.
- [18] ———, *Control of randomly varying linear dynamical systems*, Proc. Sympos. Appl. Math., 13 (1961), pp. 287–298.
- [19] H. MAASS, *Siegel's modular forms and Dirichlet series*, Lecture Notes in Math., 216, Springer-Verlag, Berlin, Heidelberg, New York, 1971.
- [20] M. MARITON, *Jump linear quadratic control with random state discontinuities*, Automatica—J. IFAC, 23 (1987), pp. 237–240.
- [21] N. E. NAHI, *Optimal recursive estimation with uncertain observation*, IEEE Trans. Inform. Theory, IT-15 (1969), pp. 457–462.
- [22] G. I. OL'SHANSKII, *Invariant cones in Lie algebras, Lie semigroups, and the holomorphic discrete series*, Functional Anal. Appl., 15 (1981), pp. 275–285.
- [23] K. R. PARTHASARATHY, *Probability measures on metric spaces*, Academic Press, New York, London, 1967.
- [24] M. A. SHAYMAN, *Phase portrait of the matrix Riccati equation*, SIAM J. Control Optim., 24 (1986), pp. 1–65.
- [25] D. L. SNYDER AND P. M. FISHMAN, *How to track a swarm of fireflies by observing their flashes*, IEEE Trans. Inform. Theory, IT-21 (1975), pp. 692–695.
- [26] C. SUNYACH, *Une classe de chaînes de Markov récurrentes sur un espace métrique complet*, Ann. Inst. H. Poincaré, 11 (1975), pp. 329–343.
- [27] A. TERRAS, *Harmonic analysis on symmetric spaces and applications II*, Springer-Verlag, Berlin, Heidelberg, New York, 1988.
- [28] M. C. VIANO, *Iterations aléatoires et filtrage de Kalman*, C.R.A.S., 305 (1987), Série 1, pp. 831–834.
- [29] P. WHITTLE, *Optimization over time*, Vol. 1, Wiley, Chichester, New York, 1982.
- [30] J. L. WILLEMS AND J. C. WILLEMS, *Robust stabilization of uncertain systems*, SIAM J. Control Optim., 21 (1983), pp. 352–374.
- [31] M. WOJTKOWSKI, *Invariant families of cones and Lyapunov exponents*, Ergodic Theory Dynamic Systems, 5 (1985), pp. 145–161.
- [32] ———, *Measure theoretic entropy of the system of hard spheres*, Ergodic Theory Dynamic Systems, 8 (1988), pp. 133–153.

A STATE-SPACE ALGORITHM FOR THE SUPEROPTIMAL HANKEL-NORM APPROXIMATION PROBLEM*

G. D. HALIKIAS†, D. J. N. LIMEBEER‡, AND K. GLOVER§

Abstract. It has been demonstrated by N. T. Young [NATO ASI Series F34, Springer-Verlag, Berlin, New York, 1987] that given a stable matrix-valued function $G_0(s)$ and a nonnegative integer k , there exists a unique superoptimal approximation $\Phi(s)$ with no more than k poles in the left half plane that minimizes the sequence $(s_1^\infty(G_0 + \Phi), s_2^\infty(G_0 + \Phi), \dots)$, with respect to lexicographic ordering, where $s_i^\infty(G_0 + \Phi) := \sup_\omega [s_i(G_0 + \Phi)(j\omega)]$ and $s_i(\cdot)$ are the singular values in descending order of magnitude. This paper presents a constructive state-space algorithm that evaluates the superoptimal approximating matrix function. The procedure recursively minimizes each frequency-dependent singular value with the aid of all-pass transformations constructed from the k th Schmidt pairs of a sequence of Hankel operators. The algorithm may be stopped after an arbitrary number of, say, $l \leq \min(m, p)$ steps. The representation formula at the l th stage will characterize all matrix functions that have $\leq k$ poles in the left half plane and that minimize $s_1^\infty(G_0 + \Phi), \dots, s_l^\infty(G_0 + \Phi)$.

Key words. model reduction, superoptimality, Hankel norm, Schmidt vectors

AMS subject classifications. 93B28, 93B40, 93B36

1. Introduction. There are many occasions on which engineers require reliable low-order approximations to high-order models. For this reason the model-order-reduction problem has been the subject of numerous theoretical investigations, and several different approaches have been developed. A technique that has received much recent attention is the optimal Hankel-norm approach [4], which offers good guaranteed performance characteristics that are close to verifiable lower bounds.

For matrix-valued problems the optimal Hankel-norm approach typically has a continuum of solutions. The question then arises as to which solution (if there is one) is best. A partial answer to this question is implicit in [4], in that \mathcal{L}^∞ -error bounds are available for only certain optimal Hankel-norm reduced-order models. Young discusses an alternative approach to the uniqueness question [16]. His suggestion is to seek to minimize the sequence $s^\infty(E) = (s_1^\infty(E), s_2^\infty(E), \dots)$ rather than just $s_1^\infty(E)$, where $s_i^\infty(E) := \sup_\omega [s_i(E)(j\omega)]$, $E(s)$ is the modeling error, and $s_i(\cdot)$ is the i th singular value (numbered in descending order of magnitude). The reduced-order model that minimizes $s^\infty(\times E)$ has been shown to exist and to be unique [16]. In model-reduction applications it is possible to reduce the \mathcal{L}^∞ -norm of the error system by using the superoptimal Hankel-norm approximation rather than some other approximation, but we have no proof of this. For diagonal problems the superoptimal solution is the most natural choice because it is the diagonal matrix of optimal solutions.

The idea behind superoptimality is easily illustrated by way of a simple 2×2 example. Suppose

$$(1.1) \quad G(s) = \begin{bmatrix} \frac{2}{s+1} & 0 \\ 0 & \frac{1}{s+1} \end{bmatrix} = \text{diag} \{g_1(s), g_2(s)\},$$

* Received by the editors January 1, 1991; accepted for publication (in revised form) January 30, 1992.

† Department of Electrical Engineering, University of Leeds, Leeds, United Kingdom.

‡ Department of Electrical Engineering, Imperial College, Exhibition Road, London, United Kingdom.

§ Department of Engineering, Trumpington Street, Cambridge, United Kingdom.

in which $\sigma_i(G) = 1, \frac{1}{2}$. Suppose also that a reduced-order approximation in $\mathcal{H}_-^\infty(1)$ is required. Any solution of the form

$$(1.2) \quad F(s) = \text{diag} \{f_1(s), -\tfrac{1}{2}\}$$

is an optimal approximation to $G(s)$, provided that $f_1(s)$ is chosen to be in $\mathcal{H}_-^\infty(1)$ such that

$$(1.3) \quad \left\| \frac{2}{s+1} + f_1 \right\|_\infty \leq \tfrac{1}{2}.$$

For example, the solution

$$(1.4) \quad F_{ap}(s) := \text{diag} \left\{ -\frac{s+3}{2(s+1)}, -\frac{1}{2} \right\}$$

results in the all-pass error system

$$(1.5) \quad (G + F_{ap})(s) = \begin{bmatrix} -\frac{1}{2} \left(\frac{s-1}{s+1} \right) & 0 \\ 0 & -\frac{1}{2} \left(\frac{s-1}{s+1} \right) \end{bmatrix},$$

and in this case $(s_1^\infty(G + F_{ap}), s_2^\infty(G + F_{ap})) = (\frac{1}{2}, \frac{1}{2})$. It is clear, however, that we can do better than this. Suppose we use our one stable pole to cancel $g_1(s)$. This gives

$$(1.6) \quad F_{so}(s) = \begin{bmatrix} -\frac{2}{s+1} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix},$$

resulting in an error system

$$(1.7) \quad (G + F_{so})(s) = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{1}{2} \left(\frac{s-1}{s+1} \right) \end{bmatrix},$$

for which $(s_1^\infty(G + F_{so}), s_2^\infty(G + F_{so})) = (\frac{1}{2}, 0)$. This is the superoptimal approximation to $G(s)$. If the $(1, 2)$ and $(2, 1)$ elements of $G(s)$ are nonzero, the situation is more complicated and a formal algorithmic procedure is required.

This paper recasts Young's algorithm in a concrete state-space framework that can be implemented on a digital computer that can tackle any rotational superoptimal approximation problem. Section 2 contains the notation to be used and a standard Hankel-norm approximation result. Section 3 contains the main results of the paper: Theorem 3.1 is standard and describes a key property of the Schmidt pairs of Hankel operators. Lemma 3.2 is a modified version of a result in [9], and Lemmas 3.3 and 3.4 are generalizations of parallel results in [9]. The main results of the paper are Theorems 3.6 and 3.6', and Algorithm 3.1, which are believed to be new. The main conclusions of our work are given in § 4.

2. Notation and preliminaries.

2.1. Notation.

$\mathbb{R}, \mathbb{R}_+, \mathbb{C}$	real, nonnegative, and complex numbers
$\mathbb{R}(s)$	field of rational functions in s with real coefficients
$\mathbb{C}_+, \bar{\mathbb{C}}_+$	open (respectively, closed) right half plane
$\mathbb{C}_-, \bar{\mathbb{C}}_-$	open (respectively, closed) left half plane
$\lambda(A), \lambda_{\max}(A)$	spectrum of a square matrix A , largest eigenvalue of A
A^*	complex conjugate transpose of $A \in \mathbb{C}^{p \times m}$
$A \geq 0, A > 0$	A is positive semidefinite (respectively, positive definite)
$A \leq 0, A < 0$	A is negative semidefinite (respectively, negative definite)
$A^\#$	generalized inverse of matrix A
$\mathcal{L}^{\infty, (p \times m)}$	space of $p \times m$ matrix functions with entries that are bounded on the $j\omega$ axis (including the point at ∞)
$\ \cdot\ _\infty$	\mathcal{L}^∞ -norm of matrices in \mathcal{L}^∞
$\mathcal{H}_+^{\infty, (p \times m)}$	subspace of \mathcal{L}^∞ ; $p \times m$ matrix functions that are analytic and bounded in \mathbb{C}_+
$\mathcal{H}_-^{\infty, (p \times m)}$	subspace of \mathcal{L}^∞ ; $p \times m$ matrix functions that are analytic and bounded in \mathbb{C}_-
$\mathcal{H}_2, \mathcal{H}_2^\perp$	the sets of functions f analytic in \mathbb{C}_+ (respectively, \mathbb{C}_-) such that $\sup_{\xi > 0} \int_{-\infty}^{\infty} \ f(\xi + j\omega)\ _2^2 d\omega < \infty$ $(\sup_{\xi < 0} \int_{-\infty}^{\infty} \ f(\xi + j\omega)\ _2^2 d\omega < \infty)$
$\mathbb{R}\mathcal{H}_2, \mathbb{R}\mathcal{H}_2^\perp$	same as $\mathcal{H}_2, \mathcal{H}_2^\perp$ except that elements are taken from $\mathbb{R}^{(p \times m)}(s)$
$\mathcal{H}_\pm^{\infty, (p \times m)}(k)$	the set of $p \times m$ matrix functions in \mathcal{L}^∞ with no more than k poles in \mathbb{C}_\pm
$\mathcal{BH}_-^\infty, \mathcal{BH}_+^\infty$	unit balls in $\mathcal{H}_-^\infty: \{f \in \mathcal{H}_-^\infty: \ f\ _\infty \leq 1\}, \{f \in \mathcal{H}_+^\infty: \ f\ _\infty \leq 1\}$
$\mathbb{R}\mathcal{L}^{\infty, (p \times m)}$	same as $\mathcal{L}^{\infty, (p \times m)}$ except that elements are taken from $\mathbb{R}^{(p \times m)}(s)$
$\mathbb{R}\mathcal{H}_\pm^{\infty, (p \times m)}(k)$	same as $\mathcal{H}_\pm^{\infty, (p \times m)}(k)$ except that elements are taken from $\mathbb{R}^{(p \times m)}(s)$
Γ_G	Hankel operator with symbol $G(s) \in \mathcal{H}_+^\infty$
$\sigma_i(G(s))$	i th Hankel singular value of $G(s)$ (i.e., of Γ_G) in decreasing order of magnitude
$[F]_+$	the stable projection of F if F is decomposed as $F := [F]_+ + [F]_-$ in which $[F]_+ \in \mathcal{H}_+^\infty$ and $[F]_- \in \mathcal{H}_-^\infty$
$s_i(A)$	i th singular value of a matrix A with the numbering in descending order (if A is a function of frequency (i.e., $A(j\omega)$), then $s_i(\cdot)$ will be a function of frequency also)
$\prod_{i=0}^n A_i$	(right) matrix product $A_0 A_1 \cdots A_n$ ($\prod_{i=0}^0 A_i := A_0$)
$\ G(s)\ _H$	$\sigma_1(G(s))$, the Hankel-norm of $G(s)$
$G^*(s)$	$G(-\bar{s})^*$, the para-Hermitian conjugate of $G(s)$
$C(A, B), \bar{C}(A, B)$	controllable and uncontrollable modes of the pair (A, B)
$O(A, C), \bar{O}(A, C)$	observable and unobservable modes of the pair (A, C)
$A \oplus B$	direct sum given by

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

Associated with a transfer function matrix $G_0(s) \in \mathbb{R}(s)^{p \times m}$ of MacMillan degree n is a state-space realization

$$G_0(s) = D + C(sI - A)^{-1}B, \quad (2.1)$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$, and $D \in \mathbb{R}^{p \times m}$. We will use the alternative notation $G_0(s) \triangleq (A, B, C, D)$ or

$$(2.2) \quad G_0(s) \triangleq \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

for realizations of $G_0(s)$. The rank of $G_0(s)$ is taken to be its rank for any s that is not a zero of $G_0(s)$.

In the above notation we have $G_0^*(s) \triangleq (-A^*, C^*, -B^*, D^*)$, and if D is nonsingular, we have $G_0^{-1}(s) \triangleq (A - BD^{-1}C, BD^{-1}, -D^{-1}C, D^{-1})$. If $G^{-1}(s) = G^*(s)$, then $G(s)$ is called *all-pass*. $G_0(s)$ is called *stable* if it has no poles in $\bar{\mathbb{C}}_+$. If $G_0(s)$ is both stable and all-pass, it is called *inner*.

We will talk about *basis changes* T in the state space of $G_0(s)$; we will take a basis change to mean $G_0(s) \triangleq (A, B, C, D) \xrightarrow{T} G_0(s) \triangleq (TAT^{-1}, TB, CT^{-1}, D)$. The *MacMillan degree* of $G_0(s)$ will be written $\deg(G_0)$, and the set of poles (zeros) of $G_0(s)$ will be denoted $\{\text{poles of } G_0\}(\{\text{zeros of } G_0\})$.

Let $P(s)$ be a partitioned matrix with a state-space realization given by

$$(2.3) \quad P(s) = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}(s) \triangleq \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{21} \end{array} \right].$$

Then

$$(2.4) \quad P_{ij}(s) = C_i(sI - A)^{-1}B_j + D_{ij}$$

is a state-space realization of $P_{ij}(s)$. A *linear fractional transformation* for the partitioned matrix P and a matrix K is defined as

$$(2.5) \quad \mathcal{F}_l(P, K) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21},$$

where K is of dimension $l \times m$ if P_{22} has dimension $m \times l$.

2.2. Preliminaries. This section provides a description of all k th-order approximations of a rational transfer function $G_0(s) \in \mathbb{R}\mathcal{H}_+^{\infty, (p \times m)}$. The description is in terms of a balanced realization of $G_0(s)$ and is based on [4, Thm. 8.7]. If $G_0(s) \triangleq (A, B, C, D)$ is balanced and minimal, the following Lyapunov equations are satisfied:

$$(2.6) \quad AP + PA^* + BB^* = 0,$$

$$(2.7) \quad A^*Q + QA + C^*C = 0,$$

in which

$$(2.8) \quad P = Q = \text{diag}(\Sigma, \sigma_{k+1}),$$

where

$$(2.9) \quad \Sigma = \text{diag}(\sigma_1, \dots, \sigma_k, \sigma_{k+2}, \dots, \sigma_n).$$

Remark 2.1. In the interests of a clear presentation, it is assumed that the $(k+1)$ th Hankel singular value is nonrepeated. Matrices A , B , and C are partitioned conformally with P and Q in (2.8) as

$$(2.10) \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = [C_1 \ C_2].$$

In a further attempt to keep the notation simple, it will be assumed that $G_0(s)$ has been scaled to give $C_2^* C_2 = B_2 B_2^* = 1$.

Lemma 2.1 parameterizes the family of all optimal Hankel-norm approximations and their corresponding extensions in terms of an (arbitrary) unstable contraction [4].

LEMMA 2.1. *Let the transfer function $G_0(s) \in \mathbb{R}\mathcal{H}_+^{\infty, (p \times m)}$ have a stable, minimal, and balanced realisation $G_0(s) \triangleq (A, B, C, D)$ with Hankel singular values $\sigma_1 \geq \dots \geq \sigma_k > \sigma_{k+1} > \sigma_{k+2} \geq \dots \geq \sigma_n > 0$, and define*

$$(2.11) \quad \Gamma = \Sigma^2 - \sigma_{k+1}^2 I_{n-1}.$$

Then all error systems $\mathcal{E}(s) = G_0(s) + \mathcal{F}(s)$ with

$$(2.12) \quad \|\mathcal{E}(s)\|_\infty = \sigma_{k+1} := \sigma$$

and $\mathcal{F}(s) \in \mathcal{H}_-^\infty(k)$ are generated by

$$(2.13) \quad \mathcal{E}(s) = \mathcal{F}_l \left(H(s), \frac{1}{\sigma} \Theta(s) \right), \quad \Theta(s) \in \mathcal{B}\mathcal{H}_-^\infty,$$

in which $H(s)H^*(s) = \sigma^2 I$ and

$$(2.14) \quad H(s) = \begin{bmatrix} G_0 + F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix},$$

where

$$(2.15) \quad F(s) = \left[\begin{array}{c|cc} \Gamma^{-1}(\sigma^2 A_{11}^* + \Sigma A_{11} \Sigma - \sigma C_1^* U B_1^*) & \Gamma^{-1}(\Sigma B_1 + \sigma C_1^* U) & -\sigma \Gamma^{-1} C_1^* C_\perp^* \\ \hline -(C_1 \Sigma + \sigma U B_1^*) & \sigma U & -\sigma C_\perp^* \\ -\sigma B_\perp B_1^* & \sigma B_\perp & 0 \end{array} \right].$$

Also, $F(s) \in \mathbb{R}\mathcal{H}_-^{\infty, (p+m-1) \times (p+m-1)}(k)$. The matrices B_\perp and C_\perp are chosen to make $[C_2 \ C_\perp^*]$ and $[B_2^* \ B_\perp^*]$ orthogonal, and U is given by $U := -C_2 B_2$. In the single-input or single-output case B_\perp or C_\perp or both will be zero and the resulting error system will be unique.

Remark 2.2. It is interesting that the construction of the superoptimal approximation is particularly simple in the special case $k = n - 1$, for which the approximation can be obtained directly from Lemma 2.1. To see this we begin by noting that all error systems are given by

$$(2.16) \quad \mathcal{E}(s) = G + F_{11} + \frac{1}{\sigma_n} F_{12} \Theta \left(I - \frac{1}{\sigma_n} F_{22} \Theta \right)^{-1} F_{21},$$

in which $F(s)$ is stable. Suppose that we set $\Theta := (1/\sigma_n) F_{12}^*(s)$, which we note has the characteristic $\|\Theta\|_\infty < 1$ by the all-pass property $H(s)H^*(s) = \sigma_n^2 I$ and since $\det(F_{12}^* F_{12}(j\omega)) \neq 0$ for all real ω . In addition, F_{22}^* is completely unstable. By exploiting the all-pass character of (2.14), the expression for the error system may be rearranged as

$$(2.17) \quad \begin{aligned} \mathcal{E}(s) &= G + F_{11} + \frac{1}{\sigma_n^2} F_{12} F_{22}^* \left(I - \frac{1}{\sigma_n^2} F_{22} F_{22}^* \right)^{-1} F_{21} \\ &= G + F_{11} + F_{12} F_{22}^* (F_{21} F_{21}^*)^{-1} F_{21} \\ &= G + F_{11} - (G + F_{11}) F_{21}^* (F_{21} F_{21}^*)^{-1} F_{21} \\ &= (G + F_{11}) (I_m - F_{21}^* (F_{21} F_{21}^*)^{-1} F_{21}). \end{aligned}$$

Since $F_{21}(j\omega)$ has rank $m-1$, the error system has rank 1 and satisfies $\|\mathcal{E}\|_\infty = \sigma_n$, which is clearly superoptimal.

To show that the error system in (2.17) is unique, we suppose there are two superoptimal approximations Φ_{so1} and Φ_{so2} . It follows from Lemma 3.4 and the unity-rank property of $\mathcal{E}(s)$ in (2.17) that there exists an all-pass matrix $W(s) = [w(s) \quad W_\perp(s)]$ such that

$$(G + \Phi_{so1})W(s) = [\sigma_n a(s)v(s) \quad 0],$$

$$(G + \Phi_{so2})W(s) = [\sigma_n a(s)v(s) \quad 0].$$

Subtracting these expressions gives

$$(\Phi_{so1} - \Phi_{so2})W(s) = 0 \Rightarrow \Phi_{so1} - \Phi_{so2}.$$

3. Main results. In this section we present the main algorithm for calculating superoptimal approximations. Following the work of Young [15], [16], the procedure is based on an inductive dimension-peeling argument. At each step of the algorithm the rank of the problem is reduced by one. Since the original problem is assumed to be of finite rank, the algorithm terminates after a finite number of steps.

The Hankel operator induced by $G_0(s)$ is defined by $\Gamma_{G_0}: \mathcal{H}_2^\perp \rightarrow \mathcal{H}_2$, $\Gamma_{G_0}g = \Pi_+ M_{G_0}g$, where Π_+ denotes the projection $\mathcal{L}_2 \rightarrow \mathcal{H}_2$ and M_{G_0} is the multiplication operator. Note that Γ_{G_0} is determined by the stable component of $G_0(s)$. We begin our development by briefly mentioning some elementary properties of the Schmidt vectors of Γ_{G_0} . The interested reader is referred to [1], [13] for a more detailed exposition. Suppose that σ_{k+1} is a singular value of Γ_{G_0} . Then there exist Schmidt vectors $f_{k+1} \in \mathcal{H}_2$ and $g_{k+1} \in \mathcal{H}_2^\perp$ that satisfy

$$(3.1) \quad \Gamma_{G_0}g_{k+1} = \sigma_{k+1}f_{k+1}$$

and

$$(3.2) \quad \Gamma_{G_0}^*f_{k+1} = \sigma_{k+1}g_{k+1},$$

and consequently

$$(3.3) \quad \Gamma_{G_0}^* \Gamma_{G_0}g_{k+1} = \sigma_{k+1}^2 g_{k+1}.$$

The next result is standard (see, e.g., [13]) and demonstrates that any Schmidt pair of Γ_{G_0} has singular-vector-type properties for the error system $(G_0 + F)$. If $F(s) \in \mathcal{H}_-^\infty(k)$ is any optimal approximation of $G_0(s) \in \mathcal{H}_+^\infty$, then

$$(3.4) \quad (G_0 + F)g_{k+1}(s) = \sigma_{k+1}f_{k+1}(s), \quad (G_0 + F)^*f_{k+1}(s) = \sigma_{k+1}g_{k+1}(s).$$

Thus by modulo scaling $f_{k+1}(s)$ and $g_{k+1}(s)$ are singular vectors of the error system $E(s) = (G_0 + F)(s)$ at each frequency $s = j\omega$ corresponding to the largest singular value of $E(j\omega)$.

THEOREM 3.1. Every $F(s) \in \mathcal{H}_-^\infty(k)$ that achieves the infimum

$$(3.5) \quad \inf_{F(s) \in \mathcal{H}_-^\infty(k)} \|G_0 + F\|_\infty = \sigma_{k+1}(G_0) < \sigma_k(G_0)$$

satisfies

$$(3.6) \quad (G_0 + F)g_{k+1}(s) = \Gamma_{G_0}g_{k+1}(s) = \sigma_{k+1}(G_0)f_{k+1}(s),$$

where (g_{k+1}, f_{k+1}) is a Schmidt pair of Γ_{G_0} corresponding to the $(k+1)$ th singular value of Γ_{G_0} .

Proof. Let $F(s)$ be any (matrix) function that achieves the infimum in (3.5), and let Γ_F be the Hankel operator that it induces. Since $F(s) \in \mathcal{H}_-^\infty(k)$, $\text{rank}(\Gamma_F) \leq k$.

Suppose (g_i, f_i) , $(i = 1, 2, \dots)$, are the Schmidt pairs of Γ_{G_0} associated with σ_i , and define P to be the orthogonal projection onto $\text{Span}(f_1, f_2, \dots, f_{k+1})$. Then $\|\Gamma_{G_0} + \Gamma_F\| = \sigma_{k+1}(G_0)$ implies that

$$(3.7) \quad \|P(\Gamma_{G_0} + \Gamma_F)\| \leq \sigma_{k+1}(G_0).$$

The operator $P\Gamma_F$ mapping $\text{Span}(g_1, g_2, \dots, g_{k+1}) \rightarrow \text{Span}(f_1, f_2, \dots, f_{k+1})$ has rank at most k , and therefore there exists a function of norm equal to one such that

$$(3.8) \quad x = \sum_{i=1}^{k+1} \alpha_i g_i \in \text{Ker}(P\Gamma_F).$$

Consequently,

$$(3.9) \quad \|P(\Gamma_{G_0} + \Gamma_F)x\|_2 = \|P\Gamma_{G_0}x\|_2 \leq \sigma_{k+1}(G_0).$$

Also,

$$(3.10) \quad \begin{aligned} \left\| P\Gamma_{G_0} \left(\sum_{i=1}^{k+1} \alpha_i g_i \right) \right\|_2 &= \left\| P \left(\sum_{i=1}^{k+1} \alpha_i \Gamma_{G_0} g_i \right) \right\|_2 = \left\| P \sum_{i=1}^{k+1} \sigma_i \alpha_i f_i \right\|_2 \\ &= \left\| \sum_{i=1}^{k+1} \sigma_i \alpha_i f_i \right\|_2 \\ &= \sqrt{\sum_{i=1}^{k+1} \sigma_i^2 |\alpha_i|^2} \end{aligned}$$

since the f_i 's are orthonormal. This implies that $\sum_{i=1}^{k+1} \sigma_i^2 |\alpha_i|^2 \leq \sigma_{k+1}^2$, and since $\sigma_{k+1} < \sigma_k \leq \dots \leq \sigma_1$ and $\sum_{i=1}^{k+1} |\alpha_i|^2 = 1$, we conclude that $\alpha_i = 0$ for $i = 1, 2, \dots, k$, so that x must be a multiple of g_{k+1} , say, $x = \beta g_{k+1}$ with $|\beta| = 1$. Now, since $x \in \text{Ker}(P\Gamma_F)$, we have that $\Gamma_F x \perp \text{Span}(f_1, f_2, \dots, f_{k+1})$ and, in particular, $\langle \Gamma_F x, f_{k+1} \rangle_2 = 0$. Consequently,

$$(3.11) \quad \begin{aligned} \|\Gamma_{G_0}x + \Gamma_Fx\|_2^2 &= \|\beta \Gamma_{G_0}g_{k+1} + \Gamma_Fx\|_2^2 \\ &= \|\sigma_{k+1}f_{k+1}\|_2^2 + \|\Gamma_Fx\|_2^2 \\ &= \sigma_{k+1}^2 + \|\Gamma_Fx\|_2^2. \end{aligned}$$

However, since $\|\Gamma_{G_0}x + \Gamma_Fx\|_2^2 \leq \|\Gamma_{G_0} + \Gamma_F\|^2 \|x\|_2^2 = \sigma_{k+1}^2$, we conclude that

$$(3.12) \quad \sigma_{k+1}^2 + \|\Gamma_Fx\|_2^2 \leq \sigma_{k+1}^2 \Rightarrow \Gamma_Fx = 0 \Rightarrow \Gamma_Fg_{k+1} = 0,$$

so that

$$(3.13) \quad (\Gamma_{G_0} + \Gamma_F)g_{k+1} = \sigma_{k+1}f_{k+1}.$$

Also, if we define Π_+ to be the stable projection operator,

$$(3.14) \quad \begin{aligned} \sigma_{k+1} &= \|\sigma_{k+1}f_{k+1}\|_2 = \|\Pi_+(G_0 + F)g_{k+1}\|_2 \\ &\leq \|(G_0 + F)g_{k+1}\|_2 \\ &\leq \|G_0 + F\|_\infty \|g_{k+1}\|_2 \\ &= \|G_0 + F\|_\infty = \sigma_{k+1}, \end{aligned}$$

since $F(s)$ achieves the infimum in (3.5). This shows that $\Pi_+(G_0 + F)g_{k+1} = (G_0 + F)g_{k+1} = \sigma_{k+1}f_{k+1}$, as required. \square

In our application we use $f_{k+1}(s)$ and $g_{k+1}(s)$ as a basis for constructing two all-pass transformations to be used in a diagonalization procedure. Lemma 3.1 represents the first step in a two-stage scaling process. The aim is to find vectors $\xi_{k+1}(s)$ and $\psi_{k+1}(s)$ that have full rank at infinity but that retain the singular vector properties of the Schmidt pair [13, Lemmas 6, 9].

LEMMA 3.1. Let $G_0(s) \triangleq (A, B, C, D)$ be the minimal and balanced realization referred to in (2.6)–(2.8). Then

$$(3.15a) \quad f_{k+1}(s) \triangleq (A, e_n, C, 0),$$

$$(3.15b) \quad g_{k+1}(s) \triangleq (-A^*, e_n, B^*, 0)$$

are a Schmidt pair for the Hankel operator Γ_{G_0} corresponding to the singular value σ_{k+1} , where $e_n = [0 \ 0 \ \cdots \ 1]^T$. Also,

$$(3.16a) \quad f_{k+1}(s) = \phi_{k+1}(s)\xi_{k+1}(s),$$

$$(3.16b) \quad g_{k+1}(s) = \phi_{k+1}^*(s)\psi_{k+1}(s),$$

in which

$$(3.17a) \quad \xi_{k+1}(s) \triangleq \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline C_1 & C_2 \end{array} \right], \quad \xi_{k+1}(j\omega) \neq 0, \quad \omega \in \mathbb{R} \cup \{\infty\},$$

$$(3.17b) \quad \psi_{k+1}(s) \triangleq \left[\begin{array}{c|c} -A_{11}^* & -A_{21}^* \\ \hline B_1^* & B_2^* \end{array} \right], \quad \psi_{k+1}(j\omega) \neq 0, \quad \omega \in \mathbb{R} \cup \{\infty\},$$

$$(3.18) \quad \phi_{k+1}(s) = \det(sI - A_{11}) / \det(sI - A).$$

Proof. For realizations (3.15a, b) see [3, p. 69]. Equations (3.16a, b) follow by invoking a result on partitioned determinants [7, p. 656]. See [9] for details.

To scale $\xi_{k+1}(s)$ to be of unit length, note that the scalar function $\xi_{k+1}^*(s)\xi_{k+1}(s)$ is positive on the imaginary axis (by 3.17a), and hence it can be spectrally factored. In particular, let $\xi_{k+1} = nd^{-1}$ be a coprime factorization over the polynomials, with $\xi_{k+1}^*\xi_{k+1} = d^{-*}n^*nd^{-1}$. Factoring $n^*n = \tilde{n}^*\tilde{n}$, where \tilde{n} is scalar with its zeros in \mathbb{C}_+ , we obtain

$$(3.19) \quad v := \xi_{k+1}(\tilde{n}d^{-1})^{-1} = n\tilde{n}^{-1} \in \mathbb{R}\mathcal{H}_-^{\infty, p \times 1},$$

with $v^*v = 1$. It is always possible to find $V_{\perp}(s)$ such that $V = [v \ V_{\perp}](s)$ is all-pass and such that $\deg(v) = \deg([v \ V_{\perp}])$. A similar argument may be used to derive a $w(s)$ from ψ_{k+1} such that $w^*w = 1$. We summarize our progress so far.

LEMMA 3.2. Given $f \in \mathbb{R}\mathcal{H}_2$ and $g \in \mathbb{R}\mathcal{H}_2^{\perp}$, there exist all-pass matrices $V \in \mathbb{R}\mathcal{H}_-^{\infty, p \times p}$ and $W \in \mathbb{R}\mathcal{H}_+^{\infty, m \times m}$ given by

$$(3.20) \quad V = [v \ V_{\perp}](s)$$

and

$$(3.21) \quad W = [w \ W_{\perp}](s),$$

in which v and w are given by (3.19) and its dual. Furthermore, minimal realizations of $V(s)$ and $W(s)$ are controllable from the first input.

Next, we give a concrete state-space construction of the vectors $v(s)$ and $w(s)$ along with their all-pass completions that are derived from standard spectral factorization theory.

Scaling $\xi(s) = C_2 + C_1(sI - A_{11})^{-1}A_{12}$ to unit length as $v(s) = \xi(s)m^{-1}(s)$ requires the solution of the spectral factorization problem $\xi^*(s)\xi(s) = m^*(s)m(s)$. Since $v(s)$ is required to be completely unstable (for reasons that will become apparent later), the spectral factor $m(s)$ must be nonminimum phase. This is achieved by choosing the appropriate solution to the corresponding Riccati equation [2]. The construction of the all-pass completion $V_{\perp}(s)$ may be achieved with no increase in degree. This accounts for the fact that minimal realizations of $V(s)$ and $W(s)$ are controllable from the first input. The construction is summarized in the following steps:

(i) If (A_{11}, A_{12}) is not completely controllable, perform a transformation T_1 in the state space of (3.17a),

$$(3.22) \quad \xi(s) \triangleq \left[\begin{array}{c|c} \frac{A_{11}}{C_1} & \frac{A_{12}}{C_2} \end{array} \right] \xrightarrow{T_1} \left[\begin{array}{c|c} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \\ \hline \tilde{C}_1 & \tilde{C}_2 \end{array} \middle| \begin{array}{c} \tilde{A}_{13} \\ 0 \\ C_2 \end{array} \right],$$

in which $(\tilde{A}_{11}, \tilde{A}_{13})$ is controllable.

(ii) Choose Ω_r as the destabilizing solution to the Riccati equation

$$(3.23) \quad \Omega_r(\tilde{A}_{11} - \tilde{A}_{13}C_2^*\tilde{C}_1) + (\tilde{A}_{11} - \tilde{A}_{13}C_2^*\tilde{C}_1)^*\Omega_r + \Omega_r\tilde{A}_{13}\tilde{A}_{13}^*\Omega_r - \tilde{C}_1^*C_\perp^*C_\perp\tilde{C}_1 = 0.$$

Since $(\tilde{A}_{11}, \tilde{A}_{13})$ is controllable by construction and since the corresponding Hamiltonian is free of $j\omega$ -axis eigenvalues ($\xi_{k+1}(j\omega)$ is full rank $\forall \omega \in \mathbb{R}$), $\Omega_r \geq 0$ is guaranteed to exist; see [8] for more details.

(iii) $V(s)$ is given by

$$(3.24) \quad V(s) = [v | V_\perp] \triangleq \left[\begin{array}{c|c} \frac{\tilde{A}_{11} + \tilde{A}_{13}(\tilde{A}_{13}^*\Omega_r - C_2^*\tilde{C}_1)}{C_\perp^*C_\perp\tilde{C}_1 + C_2\tilde{A}_{13}^*\Omega_r} & \frac{\tilde{A}_{13}}{C_2} \\ \hline \Omega_r^\# \tilde{C}_1^* C_\perp^* & C_\perp^* \end{array} \right].$$

Note that $[v | V_\perp]$ may also be described by the (nonminimal) realization

$$(3.25) \quad V(s) = [v | V_\perp] \triangleq \left[\begin{array}{c|c} \frac{A_{11} + A_{12}(A_{12}^*\Omega - C_2^*C_1)}{C_\perp^*C_\perp C_1 + C_2 A_{12}^*\Omega} & \frac{A_{12}}{C_2} \\ \hline \Omega^\# C_1^* C_\perp^* & C_\perp^* \end{array} \right],$$

in which

$$(3.26) \quad \Omega := T_1^* \begin{bmatrix} \Omega_r & 0 \\ 0 & 0 \end{bmatrix} T_1.$$

Similarly,

$$(3.27) \quad W(s) = [w | W_\perp] \triangleq \left[\begin{array}{c|c} \frac{-\bar{A}_{11}^* - \bar{A}_{31}^*(\bar{A}_{31}\tilde{\Omega}_r - B_2\bar{B}_1^*)}{B_\perp^*B_\perp\bar{B}_1^* + B_2^*\bar{A}_{31}\tilde{\Omega}_r} & \frac{-\bar{A}_{21}^*}{B_2^*} \\ \hline -\tilde{\Omega}^\# \bar{B}_1 B_\perp^* & B_\perp^* \end{array} \right],$$

where $\tilde{\Omega}_r$ denotes the destabilizing solution of

$$(3.28) \quad \tilde{\Omega}_r(\bar{A}_{11} - \bar{B}_1 B_2^* \bar{A}_{31})^* + (\bar{A}_{11} - \bar{B}_1 B_2^* \bar{A}_{31})\tilde{\Omega}_r + \tilde{\Omega}_r \bar{A}_{31}^* \bar{A}_{31} \tilde{\Omega}_r - \bar{B}_1 B_\perp^* B_\perp \bar{B}_1^* = 0$$

and the various blocks of (3.27) and (3.28) are defined by

$$(3.29) \quad \psi(s) \triangleq \left[\begin{array}{c|c} -A_{11}^* & -A_{21}^* \\ \hline B_1^* & B_2^* \end{array} \right] \xrightarrow{\tilde{T}_1} \left[\begin{array}{c|c} -\bar{A}_{11}^* & -\bar{A}_{21}^* \\ 0 & -\bar{A}_{22}^* \\ \hline \bar{B}_1^* & \bar{B}_2^* \end{array} \middle| \begin{array}{c} -\bar{A}_{31}^* \\ 0 \\ B_2^* \end{array} \right],$$

in which $(\bar{A}_{11}, \bar{A}_{31})$ is an observable pair. It is also convenient to define

$$(3.30) \quad \tilde{\Omega} = \tilde{T}_1^* \begin{bmatrix} \tilde{\Omega}_r & 0 \\ 0 & 0 \end{bmatrix} \tilde{T}_1$$

in the case that (A_{11}, A_{21}) is not completely observable.

The construction of Lemma 3.2 together with Theorem 3.1 and Lemma 3.1 imply that $V(s)$ and $W(s)$ will block-diagonalize all the optimal error systems $(G_0 + F)(s)$. Moreover, the augmented system

$$(3.31) \quad G_a(s) = \begin{bmatrix} G_0(s) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{\mathcal{H}_-^{\infty, (p+m-1) \times (p+m-1)}}$$

will have Schmidt vectors $[f_{k+1}^T(s) \ 0]^T$ and $[g_{k+1}^T(s) \ 0]^T$ and hence will be block-diagonalized by

$$(3.32) \quad V_a(s) := \begin{bmatrix} V(s) & 0 \\ 0 & I_{m-1} \end{bmatrix}, \quad W_a(s) := \begin{bmatrix} W(s) & 0 \\ 0 & I_{p-1} \end{bmatrix}.$$

The next result establishes the required decomposition of the family of all optimal error systems.

LEMMA 3.3. *The generator of all k th-order optimal error systems $G_a + F(s)$ can be diagonalized as*

$$(3.33) \quad V_a^*(G_a + F)W_a(s) = \begin{bmatrix} \sigma_{k+1}a(s) & 0 & 0 \\ 0 & G_1 + Q_{11} & Q_{12} \\ 0 & Q_{21} & Q_{22} \end{bmatrix},$$

in which

- (i) $a(s)$ is all-pass (in fact, inner),
- (ii) $G_1(s) \in \mathcal{H}_+^\infty$ and $Q(s) \in \mathcal{H}_-^\infty(k)$,
- (iii) and

$$(3.34) \quad \frac{1}{\sigma_{k+1}(G_0)} \begin{bmatrix} G_1 + Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

is all-pass.

Proof. The augmented system $G_a(s)$ has Schmidt vectors $[f_{k+1}^T(s) \ 0]^T$ and $[g_{k+1}^T(s) \ 0]^T$ corresponding to the $(k+1)$ th Hankel singular value of $G_0(s)$ and is therefore block-diagonalized by $V_a(s)$ and $W_a(s)$. The fact that $V_a(s)$, $W_a(s)$, and $(G_a + F)(s)$ are all-pass implies that

$$(3.35) \quad \frac{1}{\sigma_{k+1}(G_0)} \begin{bmatrix} G_1 + Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

and $a(s)$ are all-pass.

To show that $a(s)$ is inner, one need only show that all the poles of F_{11} (which is defined in (2.14) and (2.15)) cancel when forming $v^*(G + F_{11})w(s)$. Since we never exploit the inner character of $a(s)$ we omit the proof.

Finally, to show that the decomposition in (ii) is valid, assume that $V_a^*(s)$, $W_a(s)$, and $F(s)$ have minimal realizations

$$(3.36) \quad V_a^*(s) = \begin{bmatrix} v^* \\ V_\perp^* \\ I_{m-1} \end{bmatrix} \triangleq \left[\begin{array}{c|cc} A_v & B_v & 0 \\ C_{1v} & d_{1v} & 0 \\ C_{2v} & D_{2v} & 0 \\ 0 & 0 & I_{m-1} \end{array} \right],$$

$$(3.37) \quad W_a(s) = \begin{bmatrix} w & W_\perp \\ I_{p-1} \end{bmatrix} \triangleq \left[\begin{array}{c|cc} A_w & B_{1w} & B_{2w} & 0 \\ C_w & d_{1w} & D_{2w} & 0 \\ 0 & 0 & 0 & I_{p-1} \end{array} \right],$$

and

$$(3.38) \quad F(s) \triangleq \left[\begin{array}{c|cc} A_f & B_{1f} & B_{2f} \\ C_{1f} & D_{11f} & D_{12f} \\ C_{2f} & D_{21f} & 0 \end{array} \right]$$

with $V_a^*(s)$, $W_a(s) \in \mathbb{R}\mathcal{H}_+^\infty$, and $F(s) \in \mathbb{R}\mathcal{H}_-^\infty(k)$. Next, let $F_{12}(s)$ admit a right coprime factorization

$$(3.39) \quad F_{12}(s) := NM^{-1}(s) \stackrel{s}{=} \left[\begin{array}{c|c} A_f + B_{2f}K & B_{2f} \\ \hline C_{1f} + D_{12f}K & D_{12f} \end{array} \right] \left[\begin{array}{c|c} A_f + B_{2f}K & B_{2f} \\ \hline K & I \end{array} \right]^{-1},$$

in which K is chosen so that $\lambda(A_f + B_{2f}K) \cap \lambda(A_v) = \emptyset$; Lemma A.1 in Appendix A shows that this is always possible. Next, $v^*F_{12}(s) = 0 \Rightarrow v^*N(s) = 0$, and we may write

$$(3.40) \quad \begin{aligned} 0 &\stackrel{s}{=} \left[\begin{array}{c|c} A_v & B_v \\ \hline C_{1v} & d_{1v} \end{array} \right] \left[\begin{array}{c|c} A_f + B_{2f}K & B_{2f} \\ \hline C_{1f} + D_{12f}K & D_{12f} \end{array} \right] \\ &\stackrel{s}{=} \left[\begin{array}{cc|c} A_v & B_v(C_{1f} + D_{12f}K) & B_vD_{12f} \\ 0 & A_f + B_{2f}K & B_{2f} \\ \hline C_{1v} & d_{1v}(C_{1f} + D_{12f}K) & d_{1v}D_{12f} \end{array} \right]. \end{aligned}$$

Now let X be the (unique) solution to the linear matrix equation

$$(3.41) \quad X(A_f + B_{2f}K) - A_vX + B_v(C_{1f} + D_{12f}K) = 0.$$

Since $V_a^*(s)$ is observable through its first output, the basis change

$$(3.42) \quad T = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$$

in (3.40) establishes that

$$(3.43) \quad B_vD_{12f} + XB_{2f} = 0,$$

which when substituted back into (3.41) gives

$$(3.44) \quad XA_f - A_vX + B_vC_{1f} = 0.$$

A similar argument based on $F_{21}w(s) = 0$ will establish the existence of a matrix Y such that

$$(3.45) \quad YA_w - A_fY + B_{1f}C_w = 0$$

and

$$(3.46) \quad D_{21f}C_w - C_{2f}Y = 0.$$

Forming the product

$$(3.47) \quad \begin{aligned} &\begin{bmatrix} V_\perp^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} G + F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} W_\perp & 0 \\ 0 & I \end{bmatrix} \\ &\stackrel{s}{=} \left[\begin{array}{cccc|cc} A_v & B_vC & B_vC_{1f} & B_vD_{11f}C_w & B_vD_{11f}D_{2w} & B_vD_{12f} \\ 0 & A & 0 & BC_w & BD_{2w} & 0 \\ 0 & 0 & A_f & B_{1f}C_w & B_{1f}D_{2w} & B_{2f} \\ 0 & 0 & 0 & A_w & B_{2w} & 0 \\ \hline C_{2v} & D_{2v}C & D_{2v}C_{1f} & D_{2v}D_{11f}C_w & D_{2v}D_{11f}D_{2w} & D_{2v}D_{12f} \\ 0 & 0 & C_{2f} & D_{21f}C_w & D_{21f}D_{2w} & 0 \end{array} \right] \end{aligned}$$

and introducing the state-space transformations

$$(3.48) \quad T_1 = \begin{bmatrix} I & 0 & X & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}, \quad T_2 = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & Y \\ 0 & 0 & 0 & I \end{bmatrix}$$

together with equations (3.43)–(3.46) establishes the decomposition

$$(3.49) \quad \begin{bmatrix} V_{\perp}^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} G + F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} W_{\perp} & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} G_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},$$

in which $G_1(s) \in \mathbb{R}\mathcal{H}_+^{\infty}$ and

$$(3.50) \quad Q(s) = \left[\begin{array}{c|cc} A_f & B_{1f}D_{2w} + YB_{2w} & B_{2f} \\ \hline D_{2v}C_{1f} - C_{2v}X & 0 & D_{2v}D_{12f} \\ C_{2f} & D_{21f}D_{2w} & 0 \end{array} \right] \in \mathbb{R}\mathcal{H}_-^{\infty}(k).$$

This proves the result. \square

Remark 3.1. A direct calculation from the state-space formula for the scaled Schmidt vectors gives the following concrete realization for $Q(s)$ in terms of Ω and $\tilde{\Omega}$ defined in (3.26) and (3.30):

$$(3.51) \quad Q(s) = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \triangleq \left[\begin{array}{c|cc} \hat{A} & \hat{B}_1 & \hat{B}_2 \\ \hline \hat{C}_1 & 0 & -\sigma I \\ \hat{C}_2 & \sigma I & 0 \end{array} \right] \\ = \left[\begin{array}{c|cc} \Gamma^{-1}(\sigma^2 A_{11}^* + \Sigma A_{11} \Sigma - \sigma C_1^* U B_1^*) & (\tilde{\Omega}^{\#} + \Gamma^{-1} \Sigma) B_1 B_{\perp}^* & -\sigma \Gamma^{-1} C_1^* C_{\perp}^* \\ \hline -C_{\perp} C_1 (\Sigma + \Omega^{\#} \Gamma) & 0 & -\sigma I \\ -\sigma B_{\perp} B_1^* & \sigma I & 0 \end{array} \right].$$

Before we state and prove the main theorem of the section a technical result that gives certain properties of the zeros of the off-diagonal blocks of $Q(s)$ defined in (3.51) is required.

LEMMA 3.4. *Let $Q(s)$ be defined as in (3.51). Then*

- (i) *all MacMillan zeros of $Q_{21}(s)$ and $Q_{12}(s)$ lie in the open right half plane;*
- (ii) *if λ is a stable eigenvalue of $\hat{A} - \sigma^{-1} \hat{B}_2 \hat{C}_1$, then it is an uncontrollable mode of (\hat{A}, \hat{B}_2) , and $-\bar{\lambda}$ is a MacMillan zero of $\xi_{k+1}(s)$;*
- (iii) *if λ is a stable eigenvalue of $\hat{A} + \sigma^{-1} \hat{B}_1 \hat{C}_2$, then it is an unobservable mode of (\hat{A}, \hat{C}_2) , and λ is a MacMillan zero of $\psi_{k+1}(s)$.*

Proof. The position of the eigenvalues of $\hat{A} - \sigma^{-1} \hat{B}_2 \hat{C}_1$ will be established directly from (3.51), from which it is clear that they are located at the eigenvalues of

$$(3.52) \quad \Phi = \hat{A} + \Gamma^{-1} C_1^* C_{\perp}^* C_{\perp} C_1 (\Sigma + \Omega^{\#} \Gamma).$$

Next, we substitute (A.2) into (3.52) and make the transformation T_1 in (3.22) to obtain

$$(3.53) \quad T_1^{-*} \Gamma \Phi \Gamma^{-1} T_1^* = T_1^{-*} [C_1^* C_{\perp}^* C_{\perp} C_1 \Omega^{\#} - A_{11}^* + C_1^* C_2 A_{12}^*] T_1^* \\ = \left[\begin{array}{cc} \tilde{C}_1^* C_{\perp}^* C_{\perp} \tilde{C}_1 \Omega_r^{\#} - \tilde{A}_{11}^* + \tilde{C}_1^* C_2 \tilde{A}_{13}^* & 0 \\ * & -\tilde{A}_{22}^* \end{array} \right],$$

in which $-\tilde{A}_{22}^*$ is completely unstable and Ω_r is the appropriate solution to the reduced-dimension Riccati equation (3.23). Suppose that $\exists \xi \neq 0$ and $s_0 \in \mathbb{C}_-$ such that

$$(3.54) \quad \xi^* (\tilde{C}_1^* C_{\perp}^* C_{\perp} \tilde{C}_1 \Omega_r^{\#} - \tilde{A}_{11}^* + \tilde{C}_1^* C_2 \tilde{A}_{13}^*) = s_0 \xi^*.$$

Then the product of ξ^* , the left-hand side of (3.23), and ξ can be shown to imply that

$$(3.55) \quad (\tilde{A}_{11} - \tilde{A}_{13} C_2^* \tilde{C}_1) \xi = -s_0 \xi$$

and

$$(3.56) \quad C_{\perp} \tilde{C}_1 \xi = 0.$$

Substituting from (A.2) gives

$$(3.57) \quad \xi^*[\bar{s}_0 I + \Gamma \hat{A} \Gamma^{-1} | C_1^* C_\perp^*] = 0,$$

so that

$$(3.58) \quad (\xi^* \Gamma)[- \bar{s}_0 I - \hat{A} | \hat{B}_2] = 0.$$

This shows that $-\bar{s}_0$ is an uncontrollable mode of (\hat{A}, \hat{B}_2) . Inspecting the realization of $Q_{12}(s)$ given in (3.51), we conclude that

- (i) all the MacMillan zeros of $Q_{12}(s)$ lie in the open right half plane, and
- (ii) every eigenvalue of $\hat{A} - \sigma^{-1} \hat{B}_2 \hat{C}_1$ that lies in the left half plane is an uncontrollable mode of $[\hat{A}, \hat{B}_2]$.

Since s_0 is an unstable unobservable mode of $[\tilde{A}_{11} - \tilde{A}_{13} C_2^* \tilde{C}_1, C_\perp \tilde{C}_1]$, it is zero of

$$(3.59) \quad \begin{bmatrix} sI - \tilde{A}_{11} & -\tilde{A}_{13} \\ \tilde{C}_1 & C_2 \end{bmatrix}$$

by [12, Lemma 4.3]. Since $[\tilde{A}_{11}, \tilde{A}_{13}, \tilde{C}_1, C_2]$ is minimal, s_0 is also a MacMillan zero of $\xi(s)$.

A dual argument will establish the position of the zeros of $Q_{21}(s)$. \square

We are now in a position to present a preliminary version of the main algorithm. In this case we make a simplifying assumption about the position of the zeros of the Schmidt vectors; this restriction is removed later in Theorem 3.2'.

THEOREM 3.2. *Assume that the Schmidt vectors $\xi_{k+1}(s)$ and $\psi_{k+1}^*(s)$ given in (3.17) are free of right-half-plane MacMillan zeros. Then the family $\tilde{\mathcal{F}}(s)$ of all approximations $F \in \mathcal{H}^\infty(k)$ that minimize the pair*

$$(3.60) \quad (s_1^\infty(G_0 + F), s_2^\infty(G_0 + F))$$

lexicographically is parameterized by

$$(3.61) \quad \tilde{\mathcal{F}}(s) = F_{11} + V_\perp(-Q_{11} + \tilde{\mathcal{F}}_1)W_\perp^*(s),$$

in which $\tilde{\mathcal{F}}_1$ denotes the family of all optimal k th-order approximations of $G_1(s)$, that is, every $F_1(s) \in \mathcal{H}^\infty(k)$ that satisfies $\|G_1 + F_1\|_\infty = \sigma_{k+1}(G_1)$.

Proof. By using the results of Lemma 3.3, the family of all optimal error systems (with respect to $s_1^\infty(\cdot)$) may be parameterized by

$$(3.62) \quad G_0 + \mathcal{F}(s) = \left\{ V(s)[(\sigma_{k+1}(G_0)a(s)) \oplus (G_1 + \mathcal{F}_l(Q, \Theta))]W^*(s) \mid \Theta(s) \in \frac{1}{\sigma_{k+1}(G_0)} \mathcal{B}\mathcal{H}_-^\infty \right\}.$$

Since $V(s)$, $W^*(s)$, and $a(s)$ are all-pass, the family of all approximations that minimize $s_2(G_0 + \mathcal{F})$ are generated by

$$(3.63) \quad \inf_{\Theta \in 1/\sigma_{k+1}(G_0)\mathcal{B}\mathcal{H}_-^\infty} \|G_1 + \mathcal{F}_l(Q, \Theta)\|_\infty = s_2^\infty(G_0 + \mathcal{F}).$$

It will now be shown that if $\xi_{k+1}(s)$ and $\psi_{k+1}^*(s)$ have no MacMillan zeros in the right half plane, the minimization problem in (3.63) is equivalent to the unconstrained problem

$$(3.64) \quad \inf_{\tilde{F} \in \mathcal{H}_-^\infty(l)} \|G_1 + \tilde{F}\|_\infty = \sigma_{l+1}(G_1),$$

where l is the MacMillan degree of the stable part of $Q(s)$. Since $Q(s) \in \mathcal{H}_-^\infty(l)$ and $\|Q_{22}\Theta\|_\infty < 1$, a small-gain argument [6], [11] shows that $\mathcal{F}_l(Q, \Theta) \in \mathcal{H}_-^\infty(l)$ for all $\Theta(s) \in 1/\sigma_{k+1}(G_0)\mathcal{B}\mathcal{H}_-^\infty$. By comparing (3.63) and (3.64) we may write

$$(3.65) \quad \inf_{\Theta(s) \in 1/\sigma_{k+1}(G_0)\mathcal{B}\mathcal{H}_-^\infty} \|G_1 + \mathcal{F}_l(Q, \Theta)\|_\infty \geq \inf_{\tilde{F} \in \mathcal{H}_-^\infty(l)} \|G_1 + \tilde{F}\|_\infty = \sigma_{l+1}(G_1).$$

It will be shown that every $\tilde{F} \in \mathcal{H}_-^\infty(I)$ that achieves the minimum in (3.64) is generated by some $\Theta(s) \in 1/\sigma_{k+1}(G_0)\mathcal{B}\mathcal{H}_-^\infty$ through $\mathcal{F}_l(Q, \Theta)$. This is based on an argument given in [5].

Since all the MacMillan zeros of $Q_{12}(s)$ lie in the open right half plane (Lemma 3.4), $Q_{12}(j\omega)$ is nonsingular for every $\omega \in \mathbb{R}$. This and the all-pass character of

$$(3.66) \quad \frac{1}{\sigma_{k+1}(G_0)} \begin{bmatrix} G_1 + Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

give

$$(3.67) \quad \|G_1 + Q_{11}\|_\infty < \sigma_{k+1}(G_0).$$

Furthermore,

$$(3.68) \quad \sigma_{l+1}(G_1) \leq \|G_1 + Q_{11}\|_\infty,$$

so that

$$(3.69) \quad \sigma_{l+1}(G_1) < \sigma_{k+1}(G_0).$$

In addition, we will demonstrate that

$$(3.70) \quad \{\mathcal{F}_l(Q, \Theta): \|\Theta(s)\|_\infty < 1/\sigma_{k+1}, \Theta \in \mathcal{H}_-^\infty\}$$

generates every $\tilde{F}(s) \in \mathcal{H}_-^\infty(I)$ with the property $\|G_1 + \tilde{F}\|_\infty < \sigma_{k+1}(G_0)$. Suppose that some $\tilde{F}(s)$ satisfies $\|G_1 + \tilde{F}\|_\infty < \sigma_{k+1}(G_0)$. Then

$$(3.71) \quad \tilde{\Theta}(s) = \mathcal{F}_l\left(\frac{1}{\sigma_{k+1}^2(G_0)} \begin{bmatrix} Q_{22}^* & Q_{12}^* \\ Q_{21}^* & G_1^* + Q_{11}^* \end{bmatrix}, G_1 + \tilde{F}(s)\right)$$

gives

$$(3.72) \quad \tilde{F}(s) = \mathcal{F}_l\left(\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \tilde{\Theta}(s)\right).$$

Since $\|G_1 + Q_{11}\|_\infty < \sigma_{k+1}(G_0)$ and $\|G_1 + \tilde{F}\|_\infty < \sigma_{k+1}(G_0)$, the linear fractional map defining $\tilde{\Theta}(s)$ is well posed and, furthermore, $\|\tilde{\Theta}(s)\|_\infty < \sigma_{k+1}^{-1}(G_0)$. It remains to be shown that $\tilde{\Theta}(s) \in \mathcal{H}_-^\infty$. Suppose, contrary to what will be proved, that $\tilde{\Theta}(s)$ has r poles in the left plane. This implies that the A -matrix for $\mathcal{F}_l(Q, \tilde{\Theta})$ has $l+r$ eigenvalues in the left half plane by a Nyquist-type argument [6], [14]. Since $\tilde{F} \in \mathcal{H}_-^\infty(I)$ by assumption, there must be at least r cancellations in the closed loop. Since any cancellation is constrained to occur at a zero of (3.52a) or (3.52b) (by [10, Thm. 4.3]), we obtain the required contradiction, since the zero of (3.52a) and (3.52b) are all in the open right half plane by Lemma 3.4 and the theorem's hypothesis.

It follows, therefore, that all optimal l th order approximations of $G_1(s)$ are generated through $\mathcal{F}_l(Q, \Theta)$, as claimed. As a consequence, the set of all $F(s)$'s that minimize $(s_1^\infty(G_0 + F), s_2^\infty(G_0 + F))$ lexicographically is parameterized by

$$(3.73) \quad \begin{aligned} \tilde{\mathcal{F}}(s) &= -G_0 + V \begin{bmatrix} \sigma_{k+1}(G_0)a(s) & 0 \\ 0 & G_1(s) + \mathcal{F}(s) \end{bmatrix} W^* \\ &= -G_0 + V \begin{bmatrix} \sigma_{k+1}(G_0)a(s) & 0 \\ 0 & G_1 + Q_{11} \end{bmatrix} W^* + V \begin{bmatrix} 0 & 0 \\ 0 & \mathcal{F}_1 - Q_{11} \end{bmatrix} W^* \\ &= \mathcal{F}_l(F, 0) + V_\perp(\tilde{\mathcal{F}}_1 - Q_{11}) W_\perp^* = F_{11}(s) + V_\perp(\tilde{\mathcal{F}}_1 - Q_{11}) W_\perp^*(s), \end{aligned}$$

where $\tilde{\mathcal{F}}_1(s)$ denotes the family of all l th-order optimal approximations of $G_1(s)$. Finally, by Lemma 3.4 and the theorem's hypothesis, all stable modes of \hat{A} are

controllable through \hat{B}_2 and observable through \hat{C}_2 , and so $l = \deg(Q_+) = \deg([F_{22}]_+) = k$. This completes the proof of the theorem. \square

Remark 3.2. Theorem 3.2 establishes that all solutions that minimize $s_2^\infty(G_0 + F)$ can be parameterized in terms of all solutions that minimize $s_1^\infty(G_1 + F_1)$, which is a problem of dimension $(p-1) \times (m-1)$. The procedure can now be continued until a single-input or single-output problem is encountered (or until $\deg(G_{i+1}) \leq \deg[(Q_{i+1})_+]$). At this point there is a unique optimal approximation and the process stops.

If the assumptions on the MacMillan zeros of the Schmidt vectors $\xi_{k+1}(s)$ and $\psi_{k+1}^*(s)$ in Theorem 3.2 are relaxed, the construction of Theorem 3.2' will generate the family of superoptimal approximations with respect to the first two singular values.

THEOREM 3.2'. (i) Let $Q(s)$ be defined as in (3.51). Then its off-diagonal blocks $Q_{12}(s)$ and $Q_{21}(s)$ may be factored as $Q_{12}(s) = \mathcal{B}(s)\bar{Q}_{12}(s)$ and $Q_{21}(s) = \bar{Q}_{21}(s)\mathcal{A}(s)$, where $\mathcal{B}(s)$ and $\mathcal{A}(s)$ are inner and have degrees equal to the number of modes in $\bar{O}(\hat{A}, \hat{C}_2) \cap C(\hat{A}, \hat{B}_2) \cap \mathbb{C}_-$ and $\bar{O}(\hat{A}, \hat{C}_2) \cap \bar{C}(\hat{A}, \hat{B}_2) \cap \mathbb{C}_-$, respectively.

(ii) The family $\tilde{\mathcal{F}}(s)$ of all approximations $F \in \mathcal{H}_-^\infty(k)$ that minimize the pair

$$(3.74) \quad (s_1^\infty(G_0 + F), s_2^\infty(G_0 + F))$$

lexicographically is parameterized by

$$(3.75) \quad \tilde{\mathcal{F}}(s) = F_{11} + V_\perp \mathcal{B}(\mathcal{F}_1 - [\bar{Q}_{11}]_+ - [\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^*]_-)\mathcal{A}W_\perp^*(s),$$

in which \mathcal{F}_1 denotes the family of all optimal l th-order Hankel-norm approximations to $[\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11}]_+$, where $l = \deg([F_{22}]_+) \leq k$ and $\bar{Q}_{11} \in \mathcal{H}_-^\infty(l)$ is defined in the proof.

Proof. We begin by putting $(\hat{A}, \hat{B}_2, \hat{C}_2)$ (defined in (3.51)) into the Kalman canonical form

$$(3.76) \quad \begin{bmatrix} \hat{A} & \hat{B}_2 \\ \hat{C}_2 & 0 \end{bmatrix} = \left[\begin{array}{cccc|c} \hat{A}_{11} & 0 & \hat{A}_{13} & 0 & \hat{B}_{12} \\ \hat{A}_{21} & \hat{A}_{22} & \hat{A}_{23} & \hat{A}_{24} & \hat{B}_{22} \\ 0 & 0 & \hat{A}_{33} & 0 & 0 \\ 0 & 0 & \hat{A}_{43} & \hat{A}_{44} & 0 \\ \hline \hat{C}_{21} & 0 & \hat{C}_{23} & 0 & 0 \end{array} \right],$$

so that

$$(3.77) \quad Q(s) \triangleq \left[\begin{array}{cccc|cc} \hat{A}_{11} & 0 & \hat{A}_{13} & 0 & \hat{B}_{11} & \hat{B}_{12} \\ \hat{A}_{21} & \hat{A}_{22} & \hat{A}_{23} & \hat{A}_{24} & \hat{B}_{21} & \hat{B}_{22} \\ 0 & 0 & \hat{A}_{33} & 0 & \hat{B}_{31} & 0 \\ 0 & 0 & \hat{A}_{43} & \hat{A}_{44} & \hat{B}_{41} & 0 \\ \hline \hat{C}_{11} & \hat{C}_{12} & \hat{C}_{13} & \hat{C}_{14} & 0 & -\sigma I \\ \hat{C}_{21} & 0 & \hat{C}_{23} & 0 & \sigma I & 0 \end{array} \right].$$

Next, we transform the controllable realization $(\hat{A}_{22}, [\hat{B}_{21} \ \hat{B}_{22}], \hat{C}_{12})$ into

$$(3.78) \quad \left[\begin{array}{c|cc} \hat{A}_{22} & \hat{B}_{21} & \hat{B}_{22} \\ \hline \hat{C}_{12} & 0 & 0 \end{array} \right] \triangleq \left[\begin{array}{ccc|cc} \hat{A}_{22}^{11} & \hat{A}_{22}^{12} & \hat{A}_{22}^{13} & \hat{B}_{21}^1 & \hat{B}_{22}^1 \\ 0 & \hat{A}_{22}^u & 0 & \hat{B}_{21}^2 & \hat{B}_{22}^2 \\ 0 & \hat{A}_{22}^{32} & \hat{A}_{22}^s & \hat{B}_{21}^3 & \hat{B}_{22}^3 \\ \hline 0 & \hat{C}_{12}^2 & \hat{C}_{12}^3 & 0 & 0 \end{array} \right],$$

where the three partitions correspond to the unobservable modes (\hat{A}_{22}^{11}) and the stable and unstable unobservable modes (\hat{A}_{22}^s and \hat{A}_{22}^u , respectively). We also note that \hat{A}_{22}^{11} is completely unstable by Lemma 3.4. A similar transformation is carried out on $(\hat{A}_{33}, \hat{B}_{31}, [\hat{C}_{13}^T \ \hat{C}_{23}^T]^T)$. Combining these gives

$$(3.79) \quad Q(s) \triangleq \left[\begin{array}{cccc|cccc|cc} \hat{A}_{11} & 0 & 0 & 0 & \hat{A}_{13}^1 & \hat{A}_{13}^2 & \hat{A}_{13}^3 & 0 & \hat{B}_{11} & \hat{B}_{12} \\ \hat{A}_{21}^1 & \hat{A}_{22}^{11} & \hat{A}_{22}^{12} & \hat{A}_{22}^{13} & \hat{A}_{23}^{11} & \hat{A}_{23}^{12} & \hat{A}_{23}^{13} & \hat{A}_{24}^1 & \hat{B}_{21}^1 & \hat{B}_{22}^1 \\ \hat{A}_{21}^2 & 0 & \hat{A}_{22}^u & 0 & \hat{A}_{23}^{21} & \hat{A}_{23}^{22} & \hat{A}_{23}^{23} & \hat{A}_{24}^2 & \hat{B}_{21}^2 & \hat{B}_{22}^2 \\ \hat{A}_{21}^3 & 0 & \hat{A}_{22}^{32} & \hat{A}_{22}^s & \hat{A}_{23}^{31} & \hat{A}_{23}^{32} & \hat{A}_{23}^{33} & \hat{A}_{24}^3 & \hat{B}_{21}^3 & \hat{B}_{22}^3 \\ 0 & 0 & 0 & 0 & \hat{A}_{33}^{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{A}_{33}^{21} & \hat{A}_{33}^u & \hat{A}_{33}^{23} & 0 & \hat{B}_{31}^2 & 0 \\ 0 & 0 & 0 & 0 & \hat{A}_{33}^{31} & 0 & \hat{A}_{33}^s & 0 & \hat{B}_{31}^3 & 0 \\ 0 & 0 & 0 & 0 & \hat{A}_{43}^1 & \hat{A}_{43}^2 & \hat{A}_{43}^3 & \hat{A}_{44} & \hat{B}_{41} & 0 \\ \hline \hat{C}_{11} & 0 & \hat{C}_{12}^2 & \hat{C}_{12}^3 & \hat{C}_{13}^1 & \hat{C}_{13}^2 & \hat{C}_{13}^3 & \hat{C}_{14} & 0 & -\sigma I \\ \hat{C}_{21} & 0 & 0 & 0 & \hat{C}_{23}^1 & \hat{C}_{23}^2 & \hat{C}_{23}^3 & 0 & \sigma I & 0 \end{array} \right].$$

Next, we consider the minimal realization

$$(3.80) \quad [Q_{21} | Q_{22}] \triangleq \left[\begin{array}{ccc|cc} \hat{A}_{11} & \hat{A}_{13}^2 & \hat{A}_{13}^3 & \hat{B}_{11} & \hat{B}_{12} \\ 0 & \hat{A}_{33}^u & \hat{A}_{33}^{23} & \hat{B}_{31}^2 & 0 \\ 0 & 0 & \hat{A}_{33}^s & \hat{B}_{31}^3 & 0 \\ \hline \hat{C}_{21} & \hat{C}_{23}^2 & \hat{C}_{23}^3 & \sigma I & 0 \end{array} \right].$$

Since $\sigma^{-1}[Q_{21} \ Q_{22}]$ is part of an all-pass matrix by part (iii) of Lemma 3.3, (3.80) may be factored as

$$(3.81) \quad [Q_{21} | Q_{22}] = [\bar{Q}_{21} | Q_{22}] \begin{bmatrix} \mathcal{A} & 0 \\ 0 & I \end{bmatrix} \triangleq \left[\begin{array}{cc|cc} \hat{A}_{11} & \hat{A}_{13}^2 & \hat{B}_{11} & \hat{B}_{12} \\ 0 & \hat{A}_{33}^u & \hat{B}_{31}^2 & 0 \\ \hline \hat{C}_{21} & \hat{C}_{23}^2 & \sigma I & 0 \end{array} \right] * \left[\begin{array}{c|cc} \hat{A}_{33}^s & \hat{B}_{31}^3 & 0 \\ \hline \sigma^{-1} \hat{C}_{23}^3 & I & 0 \\ 0 & 0 & I \end{array} \right],$$

in which we may assume without loss of generality (by possibly redefining \hat{A}_{13}^3 , \hat{A}_{33}^{23} , \hat{C}_{23}^3 , \hat{B}_{11} , and \hat{B}_{31}^2) that $\mathcal{A}(s)$ is inner; details of this factorization appear in Appendix B. Similarly, a (left) inner factor $\mathcal{B}(s)$ may be extracted from $Q_{12} = \mathcal{B}(s)\bar{Q}_{12}$, leading to a realization

$$(3.82) \quad \left[\begin{array}{c} \bar{Q}_{12} \\ Q_{22} \end{array} \right] \triangleq \left[\begin{array}{cc|cc} \hat{A}_{11} & 0 & \hat{B}_{12} \\ \hat{A}_{21}^1 & \hat{A}_{22}^u & \hat{B}_{22}^2 \\ \hline \hat{C}_{11} & \hat{C}_{12}^2 & -\sigma I \\ \hat{C}_{21} & 0 & 0 \end{array} \right],$$

in which \hat{A}_{21}^1 , \hat{B}_{22}^2 , and \hat{C}_{11} may have been redefined. Next, we define

$$(3.83) \quad \bar{Q}(s) = \left[\begin{array}{cc} \bar{Q}_{11} & \bar{Q}_{12} \\ \bar{Q}_{21} & Q_{22} \end{array} \right] \triangleq \left[\begin{array}{ccc|cc} \hat{A}_{11} & 0 & \hat{A}_{13}^2 & \hat{B}_{11} & \hat{B}_{12} \\ \hat{A}_{21}^1 & \hat{A}_{22}^u & 0 & 0 & \hat{B}_{22}^2 \\ 0 & 0 & \hat{A}_{33}^u & \hat{B}_{31}^2 & 0 \\ \hline \hat{C}_{11} & \hat{C}_{12}^2 & 0 & 0 & -\sigma I \\ \hat{C}_{21} & 0 & \hat{C}_{23}^2 & \sigma I & 0 \end{array} \right].$$

Now,

$$\{\text{OHLP system zeros of the realization of } Q_{21} \text{ in (3.79)}\} \subseteq \lambda(\hat{A}_{22}^s) \cup \lambda(\hat{A}_{44}),$$

where OLHP denotes the open left half plane, implies that

$$\{\text{System zeros of the realization of } Q_{21} \text{ in (3.80)}\} \subseteq \mathbb{C}_+.$$

As a result of the all-pass factorization (3.81),

$$\begin{aligned} & \{\text{System zeros of the realization of } \bar{Q}_{21} \text{ in (3.83)}\} \\ & \subseteq \{\text{System zeros of the realization of } Q_{12}(s) \text{ in (3.80)}\} \cup \lambda(\hat{A}_{22}^u). \end{aligned}$$

We conclude that

$$\{\text{System zeros of the realization of } \bar{Q}_{21} \text{ in (3.83)}\} \subseteq \mathbb{C}_+.$$

A dual argument shows that the system zeros of the realization of \bar{Q}_{12} in (3.83) also lie in the open right half plane. Thus for any $\sigma\Theta \in \mathcal{BH}_-^\infty$

$$\begin{aligned} (3.84) \quad s_2^\infty(G_0 + \mathcal{F}) &= \|G_1 + \mathcal{F}_l(Q, \Theta)\|_\infty \\ &= \|G_1 + Q_{11} + Q_{12}\Theta(I - Q_{22}\Theta)^{-1}Q_{21}\|_\infty \\ &= \|G_1 + Q_{11} + \mathcal{B}\bar{Q}_{12}\Theta(I - Q_{22}\Theta)^{-1}\bar{Q}_{21}\mathcal{A}\|_\infty \\ &= \|\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11} + \mathcal{F}_l(\bar{Q}, \Theta)\|_\infty \\ &= \left\| [\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11}]_+ + \mathcal{F}_l\left(\begin{bmatrix} J & 0 \\ 0 & 0 \end{bmatrix} + \bar{Q}, \Theta\right) \right\|_\infty \\ &\cong \sigma_{l+1}\{[\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11}]_+\}, \end{aligned}$$

in which $J(s) := [\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11}]_-$. We may now use the arguments of Theorem 3.2 to show that

$$\mathcal{F}_l\left(\begin{bmatrix} J & 0 \\ 0 & 0 \end{bmatrix} + \bar{Q}, \Theta\right)$$

generates all σ_{k+1} -suboptimal approximations of $[\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11}]_+$ in $\mathcal{H}_-^\infty(l)$. To do this we note the following:

(i) The matrix

$$(3.85) \quad \frac{1}{\sigma} \left\{ \begin{bmatrix} \mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11} & 0 \\ 0 & 0 \end{bmatrix} + \bar{Q} \right\}$$

is all-pass.

(ii) $\|\bar{Q}_{22}\|_\infty = \|F_{22}\|_\infty < \sigma$ and the system zeros of the off-diagonal blocks of

$$(3.86) \quad \begin{bmatrix} J & 0 \\ 0 & 0 \end{bmatrix} + \bar{Q}$$

lie in the open right halfplane. This follows from the fact that $J(s)$ is completely unstable.

(iii)

$$(3.87) \quad \bar{Q} \in \mathcal{H}_-^\infty(l), \quad \text{where } l = \deg([F_{22}]_+).$$

It follows that

$$(3.88) \quad \inf_{\omega \in \mathbb{R}} \{s_2^\infty(G_0 + \mathcal{F})\} = \sigma_{l+1}([\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11}]_+).$$

Finally, the family of all approximations in $\mathcal{H}_-^\infty(k)$ that minimize the first two singular values is obtained by back substitution as

$$(3.89) \quad \tilde{\mathcal{F}}(s) = F_{11} + V_\perp \mathcal{B}(\mathcal{F}_1 - [\bar{Q}_{11}]_+ - [\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^*]_-) \mathcal{A} W_\perp^*(s),$$

where \mathcal{F}_1 denotes the family of all optimal l th-order Hankel-norm approximations to $[\mathcal{B}^*(G_1 + Q_{11})\mathcal{A}^* - \bar{Q}_{11}]_+$. \square

A simple inductive generalization of Theorem 3.2' will establish the following algorithm for calculating the (unique) superoptimal approximation.

ALGORITHM 3.1.

Given any $G_0(s) \in \mathbb{R} \mathcal{H}_+^{\infty, p \times m}$, this algorithm finds $F_{so}(s) \in \mathcal{H}_-^{\infty, p \times m}(k)$, which is the superoptimal k th-order approximation of $G_0(s)$.

1. $r = \text{rank}(G_0) \leq \min(p, m)$
2. Find $F(s)$ for $G_0(s)$ (the generator of all optimal k -th-order approximations—use state-space formulas in (2.15)); set $F_0 = F$
3. $F_{so} = F_{11}$
4. $V = I_m$ and $W = I_p$
5. For $i = 0$ to $r - 1$
 1. Find $\xi_i(s)$ and $\psi_i(s)$ corresponding to $G_i(s)$, and construct $v_i(s)$, $w_i(s)$, $V_{\perp i}(s)$, and $W_{\perp i}(s)$ using state-space formulas (3.24) and (3.27)
 2. Define $G_{i+1}(s) \in \mathcal{H}_+^\infty$ and $Q_{i+1}(s) \in \mathcal{H}_-^\infty(k)$, through decompositions given in Lemma 3.3
 3. IF $\xi_i(s)$ and/or $\psi_i^*(s)$ have right-half-plane MacMillan zeros THEN
 1. Extract all-pass common factors $\mathcal{A}(s)$ and $\mathcal{B}(s)$ (see Theorem 3.2')
 2. Redefine

$$V_{\perp i}(s) := V_{\perp i}(s)\mathcal{B}(s), \quad W_{\perp i}^*(s) := \mathcal{A}(s)W_{\perp i}^*(s),$$

$$\tilde{Q}_{i+1}(s) := [\bar{Q}_{i+1}^{11}]_+ + [\mathcal{B}^*(G_{i+1} + Q_{i+1}^{11})\mathcal{A}^*]_-,$$

$$G_{i+1}(s) := [\mathcal{B}^*(G_{i+1} + Q_{i+1}^{11})\mathcal{A}^* - \tilde{Q}_{i+1}^{11}]_+,$$
 - ELSE $\tilde{Q}_{i+1}(s) := Q_{i+1}^{11}(s)$
 4. $l = \deg([F_i^{22}]_+)$
 5. Find F_{i+1} , the generator of all l th-order optimal approximations to G_{i+1} (using (2.15))
 6. $V = VV_{\perp i}$ and $W = WW_{\perp i}$
 7. $F_{so}(s) := F_{so} + V(F_{i+1}^{11} - \tilde{Q}_{i+1})W^*(s) \in \mathcal{H}_-^\infty(k)$

Note that if the above algorithm is terminated after $l < r$ steps, the first $l + 1$ singular values will be minimized and

$$(3.90) \quad F_{so}^{(l)}(s) = F_{11}(s) + \sum_{j=0}^l \left\{ \prod_{i=0}^j V_{\perp i}(-Q_{j+1} + F_{j+1}) \left(\prod_{i=0}^j W_{\perp i} \right)^* \right\}(s)$$

will have no more than k poles in the open left half plane.

Remark 3.3. In the usual case for which the realizations of f_{k+1} and g_{k+1}^* in (3.15) are free from left-half-plane zeros, a pole-zero cancellation analysis similar to the one carried out in [9] establishes the existence of extensions that are optimal with respect to the first two singular values and are such that $\deg(F_{so}^{(2)}) \leq 2n - 3$. If this assumption holds for all G_i 's generated by Algorithm 3.1, it can be shown that

$$\deg(F_{so}) \leq \sum_{i=1}^{\text{rank}(G_0)} (n - i).$$

The cancellation analysis is moderately intricate and is consequently omitted. \square

Example 3.1. In this example we illustrate a number of the features of Algorithm 3.1. Suppose $G(s) = C(sI - A)^{-1}B$ is given by

$$(3.91) \quad A = \begin{bmatrix} -2\mu(\mu^2 - 1)^{-2} & 2(\mu^2 - 1)^{-1} & a_{13} \\ 2\mu(\mu^2 - 1)^{-1} & -1 & -1 \\ a_{31} & 2 & -2 \end{bmatrix},$$

$$B = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & b_{12} \\ \sqrt{2} & 0 \\ -3\sqrt{2}/4 & \frac{1}{2}\sqrt{7/2} \end{bmatrix}, \quad C^T = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2}\mu/(\mu^2 - 1) & -\sqrt{2}\mu/(\mu^2 - 1) \\ 1 & 1 \\ 1 & -1 \end{bmatrix},$$

in which $\mu > 1$ and

$$(3.92a) \quad a_{13} = \frac{19\mu^2 - 3 - \sqrt{7(4\mu^2 - (\mu^2 - 1)^2)}}{\sqrt{2}(\mu^2 - 1)(1 - 4\mu^2)},$$

$$(3.92b) \quad a_{31} = \frac{\sqrt{2}\mu(-1 - 3\mu^2 + \sqrt{7(4\mu^2 - (\mu^2 - 1)^2)})}{(\mu^2 - 1)(1 - 4\mu^2)},$$

$$(3.92c) \quad b_{12} = \frac{\sqrt{4\mu^2 - (\mu^2 - 1)^2}}{\mu^2 - 1}.$$

It may be verified that (3.91) is balanced with controllability and observability gram-mians

$$(3.93) \quad P = Q = \Sigma = \text{diag}(\mu/2, 0.5, 0.25).$$

By applying the triangularizing transformation

$$(3.94) \quad T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2\mu(\mu^2 - 1)^{-1} & 0 & 1 \end{bmatrix}$$

to

$$(3.95) \quad \begin{bmatrix} sI - A_{11} & -A_{12} \\ C_1 & C_2 \end{bmatrix},$$

we conclude that the scaled Schmidt vector $\xi_3(s) = C_2 + C_1(sI - A_{11})^{-1}A_{12}$ corresponding to $\sigma_3 = 0.25$ has a single MacMillan zero in the open right half plane for every $\mu > 1$. In the remainder of this example we fix μ at $\mu = 1.2$ and seek the superoptimal Hankel-norm approximation to $G(s)$ over $\mathcal{H}_-^\infty(2)$.

A realization of $F(s)$ (defined in (2.15)) that generates all optimal approximations with respect to the first singular value may be calculated as

$$(3.96) \quad F(s) \triangleq \left[\begin{array}{cc|ccc} -11.196 & 0 & 3.8569 & 5.5031 & 0 \\ -5.0997 & -1.6667 & 2.6667 & 0 & -1.3333 \\ \hline -1.2868 & -0.4861 & 0.1326 & -0.1169 & -0.1768 \\ 1.2868 & -0.2210 & -0.1326 & 0.1169 & -0.1768 \\ \hline -0.8278 & -0.1654 & 0.1654 & 0.1875 & 0 \end{array} \right],$$

in which we note that a stable mode ($\lambda = -11.196$) is uncontrollable through the last column of the B -matrix (\hat{B}_2). The decomposition of Lemma 3.3 is now carried out to

give

$$(3.97) \quad Q(s) \triangleq \left[\begin{array}{cc|cc} -11.196 & 0 & 8.0841 & 0 \\ -5.0997 & -1.6667 & -1.9608 & -1.3333 \\ \hline 0 & -0.5767 & 0 & -0.2500 \\ -0.8278 & -0.1654 & 0.2500 & 0 \end{array} \right]$$

and

$$(3.98) \quad G_1(s) \triangleq \left[\begin{array}{cc|c} -0.9485 & -2.6797 & -2.6388 \\ 1.7808 & -11.662 & 4.7682 \\ \hline 0.7585 & 0.13606 & 0 \end{array} \right],$$

which may be written in transfer function form as

$$(3.99) \quad Q(s) = \left[\begin{array}{cc} \frac{1.3727(s+32.221)}{(s+11.196)(s+1.6667)} & -\frac{0.25(s-1.4142)}{s+1.6667} \\ \frac{0.25(s-1.4142)(s-11.196)}{(s+1.6667)(s+11.196)} & \frac{0.22048}{s+1.6667} \end{array} \right]$$

and

$$(3.100) \quad G_1(s) = -\frac{1.3532(s+24.434)}{(s+11.196)(s+1.4142)}.$$

Note that because of the presence of a minimum-phase system zero in the realization of $Q_{12}(s)$ in (3.97) (due to the uncontrollable mode $\lambda = -11.198$), $\mathcal{F}_l(Q, \sigma_3^{-1}\mathcal{B}\mathcal{H}^\infty)$ does not generate all the $\sigma_3 = 0.25$ -suboptimal approximations of $G_1(s)$ in $\mathcal{H}^\infty(2)$. We may now extract the inner factor

$$(3.101) \quad \mathcal{A}(s) = \frac{s-11.196}{s+11.196}$$

corresponding to this zero, so that

$$(3.102) \quad G_1(s) + \mathcal{F}_l(Q, \sigma_3^{-1}\Theta) = (\bar{G}_1 + \mathcal{F}_l(\bar{Q}, \sigma_3^{-1}\Theta))\mathcal{A}(s),$$

in which

$$(3.103) \quad \bar{G}_1 := -\frac{1.3532(s+24.434)}{(s-11.196)(s+1.4142)}$$

and

$$(3.104) \quad \bar{Q}(s) = \left[\begin{array}{cc} \frac{1.3727(s+32.221)}{(s-11.196)(s+1.6667)} & -\frac{0.25(s-1.4142)}{s+1.6667} \\ \frac{0.25(s-1.4142)}{s+1.6667} & \frac{0.22048}{s+1.6667} \end{array} \right].$$

Furthermore, it may be verified that

$$(3.105) \quad \mathcal{F}_l\left(\bar{Q} + \begin{bmatrix} [\bar{G}_1]_- & 0 \\ 0 & 0 \end{bmatrix}, \sigma_3^{-1}\mathcal{B}\mathcal{H}^\infty\right)$$

generates all σ_3 -suboptimal approximations of $[\bar{G}_1]_+$ in $\mathcal{H}^\infty(1)$, so that

$$(3.106) \quad s_2^\infty(G+F) = \inf_{F_1 \in \mathcal{H}^\infty(1)} \|[\bar{G}_1]_+ + F_1\|_\infty = 0,$$

which is in agreement with Remark 2.2. The superoptimal approximation is finally obtained as

$$(3.107) \quad F_{so} = F_{11} - V_{\perp}(G_1 + Q_{11})W_{\perp}^* \\ = \left[\begin{array}{ccc|cc} -11.244 & 0.5310 & 0 & 4.0836 & 5.6906 \\ -8.8849 & -1.3664 & 0 & 0.9722 & -1.2665 \\ 0 & 0 & 1.4142 & -0.1467 & -0.0541 \\ \hline -1.3554 & -0.6586 & -0.1297 & 0.1326 & -0.1169 \\ 1.1951 & -0.3677 & 0.3130 & -0.1326 & 0.1169 \end{array} \right].$$

4. Conclusions. The purpose of this paper is to develop an implementable state-space version of Young's algorithm for superoptimal Hankel-norm approximations [16]. The new Algorithm 3.1 requires only standard linear algebraic library routines and has the added advantage that it can be stopped after only $l < r$ steps. In this event, the procedure produces a representation formula for all the approximations in $\mathcal{H}_{-}^{\infty}(k)$ that minimize $\{s_1^{\infty}(G_0 + \mathcal{F}), s_2^{\infty}(G_0 + \mathcal{F}), \dots, s_l^{\infty}(G_0 + \mathcal{F})\}$ with respect to lexicographic ordering.

Appendix A.

LEMMA A.1. *Let \tilde{A} denote the A -matrix of a minimal realization of $V^*(s)$. Then there exists a matrix K such that $\lambda(A_f + B_{2f}K) \cap \lambda(\tilde{A}) = \emptyset$.*

Proof. Since \tilde{A} is stable, it suffices to show that every stable eigenvalue of A_f that is uncontrollable through B_{2f} does not belong to the spectrum of \tilde{A} . Suppose for contradiction that λ is such an eigenvalue. Then $\exists \beta \neq 0$ such that

$$(A.1) \quad \beta^*[\lambda I - A_f | B_{2f}] = 0.$$

It follows by using (2.15), (2.6), and (2.7) that

$$(A.2) \quad \Gamma^{-1}A_f^*\Gamma + \Sigma\Gamma^{-1}C_1^*C_{\perp}^*C_1 = \Gamma^{-1}(\sigma^2A_{11} + \Sigma A_{11}^*\Sigma - \sigma B_1U^*C_1)\Gamma \\ + \Sigma\Gamma^{-1}C_1^*(I - C_2C_2^*)C_1 \\ = \Gamma^{-1}(\sigma^2A_{11} + \sigma B_1B_2^*C_2^*C_1 - \Sigma^2A_{11} - \Sigma C_1^*C_2C_2^*C_1) \\ = -A_{11} + A_{12}C_2^*C_1.$$

Substituting (A.2) into (A.1) yields

$$(A.3) \quad \tilde{\beta}^*[\lambda I + A_{11}^* + C_1^*C_2A_{12}^* | C_1^*C_{\perp}^*] = 0,$$

where $\tilde{\beta} = \Gamma^{-1}\beta$. Next, we introduce the transformation T_1 , defined in (3.22), to (A.3). This gives

$$(A.4) \quad \begin{bmatrix} \hat{\beta}_1^* & \hat{\beta}_2^* \end{bmatrix} \begin{bmatrix} \lambda I + \tilde{A}_{11}^* - \tilde{C}_1^*C_2\tilde{A}_{13}^* & 0 & \tilde{C}_1^*C_{\perp}^* \\ \tilde{A}_{12}^* - \tilde{C}_2^*C_2\tilde{A}_{13}^* & \lambda I + \tilde{A}_{22}^* & \tilde{C}_2^*C_{\perp}^* \end{bmatrix} = 0,$$

which implies that

$$(A.5) \quad \hat{\beta}_1^*(\lambda I + \tilde{A}_{11}^* - \tilde{C}_1^*C_2\tilde{A}_{13}^*) + \hat{\beta}_2^*(\tilde{A}_{12}^* - \tilde{C}_2^*C_2\tilde{A}_{13}^*) = 0,$$

$$(A.6) \quad \hat{\beta}_2^*(\lambda I + \tilde{A}_{22}^*) = 0,$$

$$(A.7) \quad \hat{\beta}_1^*\tilde{C}_1^*C_{\perp}^* + \hat{\beta}_2^*\tilde{C}_2^*C_{\perp}^* = 0.$$

Now, since $\lambda \in \mathbb{C}_{-}$ and $\lambda(-\tilde{A}_{22}^*) \subseteq \lambda(-A_{11}^*) \subseteq \mathbb{C}_{+}$,

$$(A.8) \quad (A.6) \Rightarrow \hat{\beta}_2^* = 0 \Rightarrow \hat{\beta}_1^* \neq 0.$$

Thus

$$(A.9) \quad (A.5) \Rightarrow \hat{\beta}_1^*(\lambda I + \tilde{A}_{11}^* - \tilde{C}_1^* C_2 \tilde{A}_{13}^*) = 0,$$

$$(A.10) \quad (A.7) \Rightarrow \hat{\beta}_1^* \tilde{C}_1^* C_\perp^* = 0.$$

From the multiplication of $\hat{\beta}_1^*$, the left-hand side of (3.23), and $\hat{\beta}_1$ we obtain

$$(A.11) \quad -2 \operatorname{Re}(\lambda) \hat{\beta}_1^* \Omega_r \hat{\beta}_1 + \hat{\beta}_1^* \Omega_r \tilde{A}_{13} \tilde{A}_{13}^* \Omega_r \hat{\beta}_1 = 0 \Rightarrow \Omega_r \hat{\beta}_1 = 0,$$

since $\operatorname{Re}(\lambda) < 0$. It follows from (3.24) that

$$(A.12) \quad v^*(s) \triangleq \left[\frac{-\tilde{A}_{11}^* - (\Omega_r \tilde{A}_{13} - \tilde{C}_1^* C_2) \tilde{A}_{13}^*}{-\tilde{A}_{13}^*} \middle| \frac{\tilde{C}_1^* C_\perp^* C_\perp + \Omega_r \tilde{A}_{13} C_2}{C_2^*} \right],$$

and it is easy to show by using (A.11) that λ is an uncontrollable mode of this realization, which proves the result. \square

Appendix B.

LEMMA B.1. $[Q_{21} | Q_{22}]$ in (3.80) may be factored as

$$(B.1) \quad \left[\begin{array}{cc|cc} \hat{A}_{11} & \hat{A}_{13}^2 & \hat{B}_{11} & \hat{B}_{12} \\ 0 & \hat{A}_{33}^u & \hat{B}_{31}^2 & 0 \\ \hline \hat{C}_{21} & \hat{C}_{23}^2 & \sigma I & 0 \end{array} \right] * \left[\begin{array}{cc|cc} \hat{A}_{33}^s & \hat{B}_{31}^3 & 0 \\ \hline \sigma^{-1} \hat{C}_{23}^3 & I & 0 \\ 0 & 0 & I \end{array} \right],$$

in which $\mathcal{A}(s) \triangleq (\hat{A}_{33}^s, \hat{B}_{31}^3, \sigma^{-1} \hat{C}_{23}^3, I)$ is inner.

Proof. Consider the minimal realization

$$(B.2) \quad [Q_{21} | Q_{22}] \triangleq \left[\begin{array}{ccc|cc} \hat{A}_{11} & \hat{A}_{13}^2 & \hat{A}_{13}^3 & \hat{B}_{11} & \hat{B}_{12} \\ 0 & \hat{A}_{33}^u & \hat{A}_{33}^{23} & \hat{B}_{31}^2 & 0 \\ 0 & 0 & \hat{A}_{33}^s & \hat{B}_{31}^3 & 0 \\ \hline \hat{C}_{21} & \hat{C}_{23}^2 & \hat{C}_{23}^3 & \sigma I & 0 \end{array} \right]$$

with controllability grammian

$$(B.3) \quad P = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^* & P_{22} \end{bmatrix},$$

in which $\dim(P_{11}) = \dim(\hat{A}_{11}) + \dim(\hat{A}_{33}^u)$. We will assume the following without loss of generality:

(i) $P_{12} = 0$. This may be achieved by introducing the state-space transformation

$$(B.4) \quad T = \begin{bmatrix} I & -P_{12}P_{22}^{-1} \\ 0 & I \end{bmatrix}$$

in (B.2) while noting that \hat{A}_{33}^s asymptotically stable implies that $P_{22} > 0$. Note also that $\hat{A}_{13}^3, \hat{A}_{33}^{23}, \hat{C}_{23}^3, \hat{B}_{11}$, and \hat{B}_{31}^2 are redefined as the result of this transformation.

(ii) $P_{22} = I$, i.e., the realization $(\hat{A}_{33}^s, \hat{B}_{31}^3)$ is input balanced.

Next, consider the all-pass equations corresponding to

$$(B.5) \quad \sigma^{-1} [Q_{21} | Q_{22}] \triangleq \left[\begin{array}{ccc|cc} \hat{A}_{11} & \hat{A}_{13}^2 & \hat{A}_{13}^3 & \hat{B}_{11} & \hat{B}_{12} \\ 0 & \hat{A}_{33}^u & \hat{A}_{33}^{23} & \hat{B}_{31}^2 & 0 \\ 0 & 0 & \hat{A}_{33}^s & \hat{B}_{31}^3 & 0 \\ \hline \sigma^{-1} \hat{C}_{21} & \sigma^{-1} \hat{C}_{23}^2 & \sigma^{-1} \hat{C}_{23}^3 & I & 0 \end{array} \right] \\ \triangleq \left[\begin{array}{cc|cc} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} & \tilde{B}_{12} \\ 0 & \hat{A}_{33}^s & \hat{B}_{31}^3 & 0 \\ \hline \sigma^{-1} \tilde{C}_1 & \sigma^{-1} \hat{C}_{23}^3 & I & 0 \end{array} \right].$$

These may be written out in full as

$$(B.6) \quad \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \hat{A}_{33}^s \end{bmatrix} \begin{bmatrix} P_{11} & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} P_{11} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \hat{A}_{33}^s \end{bmatrix}^* \\ + \begin{bmatrix} \tilde{B}_{11} & \tilde{B}_{12} \\ \hat{B}_{31}^3 & 0 \end{bmatrix} \begin{bmatrix} \tilde{B}_{11} & \tilde{B}_{12} \\ \hat{B}_{31}^3 & 0 \end{bmatrix}^* = 0$$

and

$$(B.7) \quad [I \quad 0] \begin{bmatrix} \tilde{B}_{11} & \tilde{B}_{12} \\ \hat{B}_{31}^3 & 0 \end{bmatrix} + [\sigma^{-1} \tilde{C}_1 \quad \sigma^{-1} C_{23}^3] \begin{bmatrix} P_{11} & 0 \\ 0 & I \end{bmatrix} = 0,$$

from which we get

$$(B.8) \quad (\hat{B}_{31}^3)^* = -\sigma^{-1} \hat{C}_{23}^3$$

and

$$(B.9) \quad \tilde{A}_{12} = -\sigma^{-1} \tilde{B}_{11} \hat{C}_{23}^3.$$

(B.8) and (B.9) may now be used to establish the required decomposition, while (B.8) and the (2, 2) block of (B.6) show that $\mathcal{A}(s)$ is inner. \square

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31-73.
- [2] B. D. O. ANDERSON, *An algebraic solution to the spectral factorization problem*, IEEE Trans. Automat. Control., AC-12 (1967), pp. 410-414.
- [3] B. A. FRANCIS, *A Course in \mathcal{H}^∞ Control Theory*, Springer-Verlag, Berlin, New York, 1987.
- [4] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariate systems and their \mathcal{L}^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.
- [5] —, *Model reduction: A tutorial on Hankel-norm methods and lower bounds on L^2 errors*, Proc. 10th Triennial International Federation of Automatic Control World Congress, Pergamon Press, Munich, 10 (1987), pp. 288-293.
- [6] K. GLOVER, R. CURTAIN, AND J. PARTINGTON, *Realization and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863-898.
- [7] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [8] V. KUCERA, *A contribution to matrix Riccati equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344-347.
- [9] D. J. N. LIMEBEER, G. D. HALIKIAS, AND K. GLOVER, *A state-space algorithm for the computation of super-optimal matrix interpolating functions*, Internat. J. Control, 50 (1989), pp. 2431-2466.
- [10] D. J. N. LIMEBEER AND Y. HUNG, *An analysis of the pole-zero cancellations in \mathcal{H}^∞ -optimal control problems of the first kind*, SIAM J. Control Optim., 25 (1987), pp. 1457-1493.
- [11] D. J. N. LIMEBEER AND B. D. O. ANDERSON, *An interpolation theory approach to \mathcal{H}^∞ controller degree bounds*, Linear Algebra Appl., 98 (1988), pp. 347-386.
- [12] D. J. N. LIMEBEER AND G. D. HALIKIAS, *An analysis of the pole-zero cancellations in \mathcal{H}^∞ -optimal control problems of the second kind*, SIAM J. Control Optim., 26 (1988), pp. 646-677.
- [13] J. R. PARTINGTON, *An Introduction to Hankel Operators*, Cambridge University Press, London, 1988.
- [14] R. M. REDHEFFER, *On a certain linear fractional transformation*, J. Math. Phys., 39 (1960), pp. 269-286.
- [15] N. J. YOUNG, *The Nevanlinna-Pick problem for matrix-valued functions*, J. Operator Theory, 15 (1986), pp. 239-265.
- [16] —, *Super-optimal Hankel norm approximations*, in Modelling, Robustness and Sensitivity Reduction in Control Systems, NATO ASI Series F34, Springer-Verlag, Berlin, New York, 1987.

DYNAMICAL BOUNDARY CONTROL FOR ELASTIC PLATES OF GENERAL SHAPE*

LAWRENCE MARKUS† AND YUNCHENG YOU‡

Abstract. The control of transverse vibrations of elastic plates of general shape by feedback boundary control is formulated as an abstract evolution equation. Because the control acts locally on the boundary, which possesses a flanged rim with inertial properties of mass and bending moment, the analysis concerns dynamical controllability and stabilizability of a hybrid system. By the approach of energy decay inequalities and Hörmander's global uniqueness theorem, it is shown that the system is strongly stabilizable by a locally supported damping feedback of boundary velocity and boundary angular velocity, and hence the system is approximately controllable.

Key words. dynamical boundary control, hybrid system, elastic plate, abstract evolution equation, stabilization

AMS subject classifications. 35Q72, 47D06, 73C02, 93C20

1. Geometry of hybrid control systems: summary of results. Earlier investigations [10]–[12] established the ability to control and stabilize the transverse vibrations of an elastic beam, clamped at one end, by feedback damping with dynamical control forces and torques acting on the other end. The present study was motivated by the control requirements of a space satellite supporting an elastic beam (governed by the Euler-Bernoulli partial differential equation) attached to an antenna with mass and moment of inertia (governed by Newtonian ordinary differential equations) on which the control acts by means of rocket thrust and couples. In this sense the total system is a hybrid system consisting of a partial differential equation and two coupled ordinary differential equations through which the control dynamics are filtered.

This type of hybrid control system was generalized [17], [18] to two-dimensional, rectangular elastic plates with inertial properties along the controlled edge, which is rimmed with a flange or lip that has inertial properties. Here the distributed system is governed by the Petrovsky equation for the small displacement $u(t, x, y)$ of a homogeneous, thin elastic plate over the region Ω at time $t \geq 0$,

$$\frac{\partial^2 u}{\partial t^2} + \Delta^2 u = 0 \quad \text{in } R^+ \times \Omega,$$

with suitable boundary conditions at the three clamped sides and dynamical boundary control from the free side. As usual, Δ^2 is the biharmonic operator and the physical units are appropriately chosen. The control system is then reformulated as an abstract evolution equation in a suitable Hilbert space. This framework will also be used in the present investigation.

It is important to note that in all these previous investigations of elastic beams and rectangular plates with dynamical boundary control, the explicit distribution of eigenvalues plays an essential role. The methods in [11] show that the stabilization of beam vibration proceeds with a subexponential decay rate. It is possible, and this is an open issue, that this behavior is a feature common to all similar hybrid systems.

* Received by the editors August 26, 1991; accepted for publication (in revised form) February 25, 1992.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. This author was supported in part by National Science Foundation grant DMS 90-02919.

‡ Department of Mathematics, University of South Florida, Tampa, Florida 33620. This author was supported in part by the University of South Florida Research and Creative Scholarship grant 1249-930-RO.

Of course, there is a highly developed theory of the exact controllability and stabilization of elastic systems, including plates governed by the usual (static) boundary control without any inertial properties arising from the boundary (see [3]–[9]). Since these other configurations have no dynamical equations ascribed to the controlled boundary, they do not fit the theory of hybrid control systems as it is developed in this paper. To the authors' knowledge, the well-known Hilbert uniqueness method [9] does not apply directly to our problem of the strong stabilization of hybrid control systems, because if the Hilbert uniqueness method could be adapted to this problem, then it would result in exact controllability and stabilization at a uniform exponential rate. However, by such a dynamical boundary-damping feedback one generally cannot expect an exponential decay rate, for some intrinsic reasons, as is indicated in [11] for the case of a one-dimensional beam.

Here we consider the small transverse deflection $u(t, x, y)$ of a thin, isotropic, homogeneous elastic plate over a region Ω at time $t \geq 0$. The region Ω is an open, bounded, and connected set in R^2 , with a boundary $\Gamma = \partial\Omega$ consisting of a finite set of simple closed curves. Assume that Γ is piecewise smooth and that the inward uniform cone condition is satisfied everywhere on Γ . For easy reference we shall designate Ω as a *region of finite genus*.

Assume that the boundary $\Gamma = \partial\Omega$ consists of two disjoint pieces: Γ_0 on which the plate is clamped and Γ_1 along which the plate is free on part Γ_f of Γ_1 and is controlled on $\Gamma_c = \Gamma_1 \setminus \Gamma_f$. Here, Γ_0 (the union of arcs on some boundary curves of Ω) is a closed subset of Γ with a nonempty relative interior in Γ , so Γ_1 is relatively open in Γ . Also, the controlled part Γ_c is relatively open in Γ_1 , and its linear measure is positive, i.e., $\text{meas}_1(\Gamma_c) > 0$. Moreover, the inertial properties of the rim are restricted to Γ_c only.

Following the classical elasticity theory [3]–[5], [14] of small vibrations $u(t, x, y)$ of thin, isotropic, homogeneous elastic plates over a plane region Ω of finite genus, we can formulate the aforementioned boundary control problem as follows:

$$\begin{aligned}
 & \frac{\partial^2 u}{\partial t^2} + \Delta^2 u = 0 && \text{in } R^+ \times \Omega, \\
 & u = \frac{\partial u}{\partial n} = 0, && \text{on } R^+ \times \Gamma_0 \text{ (clamped boundary } \Gamma_0), \\
 & \Delta u + (1 - \sigma)A_2 u = 0 && \text{on } R^+ \times \Gamma_f \text{ (no shearing force on } \Gamma_f), \\
 (1) \quad & \frac{\partial \Delta u}{\partial n} + (1 - \sigma)A_1 u = 0 && \text{on } R^+ \times \Gamma_f \text{ (no bending torque on } \Gamma_f), \\
 & \frac{\partial^2 u}{\partial t^2} = k_1 \left[\frac{\partial \Delta u}{\partial n} + (1 - \sigma)A_1 u \right] + f_1(t, x, y) && \text{on } R^+ \times \Gamma_c \text{ (control force),} \\
 & \frac{\partial^2 u_n}{\partial t^2} = -k_2 [\Delta u + (1 - \sigma)A_2 u] + f_2(t, x, y) && \text{on } R^+ \times \Gamma_c \text{ (control torque),} \\
 & u(0, x, y) = u_0(x, y) \quad \text{and} \quad u_t(0, x, y) = u_1(x, y) && \text{in } \Omega \text{ (initial data).}
 \end{aligned}$$

Here the inertial properties of the boundary are supported along the rim of Γ_c where the control shearing force (per unit boundary mass) $f_1(t, x, y)$ and the bending torque (without twisting) $f_2(t, x, y)$ act as dynamic controllers. Here $k_1^{-1} = \rho > 0$ is the linear boundary density and $k_2^{-1} = J > 0$ is the bending moment of inertia per unit length of the boundary, and these are constants with appropriate physical units.

The two operators A_1 and A_2 are defined by

$$(2) \quad A_1 u = \frac{\partial}{\partial s} [(n_1^2 - n_2^2)u_{xy} + n_1 n_2 (u_{yy} - u_{xx})], \quad A_2 u = 2n_1 n_2 u_{xy} - n_1^2 u_{yy} - n_2^2 u_{xx},$$

where $n = (n_1, n_2)$ is the outward unit normal vector to Γ and s is the arc length sensed by $(-n_2, n_1)$ along Γ . The physical constant $0 < \sigma < \frac{1}{2}$ is the Poisson ratio of elasticity.

Example. Let Ω be an annulus described in polar coordinates (r, θ) by

$$\Omega = \{(r, \theta) : 0 < \delta < r < 1, 0 \leq \theta < 2\pi\},$$

with Γ_0 being the inner circle ($r = \delta$) and $\Gamma_c = \Gamma_1$ being the outer circle ($r = 1$). Assume that the elastic plate over Ω is clamped on the inner circle Γ_0 and that the boundary dynamic controls act on Γ_c . We have $r = 1$, $s = \theta$, and $(n_1, n_2) = (\cos \theta, \sin \theta)$, so $u_n = \partial u / \partial r$ and the operators A_1 and A_2 take the form

$$A_1 u = \frac{\partial}{\partial \theta} \left[\cos(2\theta) u_{xy} + \frac{1}{2} \sin(2\theta) (u_{yy} - u_{xx}) \right],$$

$$A_2 u = \sin(2\theta) u_{xy} - \cos^2 \theta u_{yy} - \sin^2 \theta u_{xx}.$$

This example will be used later to illustrate our general results of Theorems 1 and 2.

Our main results in this paper assert that for the hybrid control system of an elastic plate of general shape (Ω is a plane region of finite genus), the corresponding evolution system is strongly stabilizable by boundary-damping feedback

$$f_1 = -\frac{\partial u}{\partial t}(t, x, y) \quad \text{and} \quad f_2 = -\frac{\partial^2 u}{\partial t \partial n}(t, x, y) \quad \text{on } R^+ \times \Gamma_c.$$

Consequently, this hybrid system is approximately controllable by open-loop controllers $f_1(t, x, y)$ and $f_2(t, x, y)$ supported on the arc Γ_c .

2. Abstract evolution equation. In this section we set a framework to deal with the hybrid control system described by (1). Denote by H the product Hilbert space

$$H = L^2(\Omega) \times L^2(\Gamma_c) \times L^2(\Gamma_c),$$

with the inner product defined by

$$\langle (u, u_1, u_2), (v, v_1, v_2) \rangle = \iint_{\Omega} uv \, dx \, dy + \frac{1}{k_1} \int_{\Gamma_c} u_1 v_1 \, ds + \frac{1}{k_2} \int_{\Gamma_c} u_2 v_2 \, ds.$$

Then H is a separable real Hilbert space. Here Ω is a plane region of finite genus with boundary $\Gamma = \Gamma_0 \cup \Gamma_f \cup \Gamma_c$ as specified above. Define operators A and B by

$$(3) \quad A(u, u_1, u_2) = \begin{bmatrix} \Delta^2 u & 0 & 0 \\ -k_1(\partial \Delta u / \partial n + (1 - \sigma)A_1 u)|_{\Gamma_c} & 0 & 0 \\ k_2(\Delta u + (1 - \sigma)A_2 u)|_{\Gamma_c} & 0 & 0 \end{bmatrix},$$

with the domain

$$D(A) = \left\{ (u(x, y), u_1(s), u_2(s)) \in H^4(\Omega) \times L^2(\Gamma_c) \times L^2(\Gamma_c) : u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma_0, \right. \\ \left. \Delta u + (1 - \sigma)A_2 u = 0, \quad \frac{\partial \Delta u}{\partial n} + (1 - \sigma)A_1 u = 0 \quad \text{on } \Gamma_f, \quad u = u_1, \quad \frac{\partial u}{\partial n} = u_2 \quad \text{on } \Gamma_c \right\},$$

and

$$(4) \quad B = \begin{pmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{pmatrix},$$

with I being the identity on $L^2(\Gamma_c)$. Thus we have $A: D(A) \rightarrow H$, and B is a bounded linear operator from $L^2(\Gamma_c) \times L^2(\Gamma_c)$ to H . Denote by

$$v(t) = \begin{bmatrix} u(t, x, y) | (x, y) \in \Omega \\ u(t, x, y) | (x, y) \in \Gamma_c \\ (\partial u / \partial n)(t, x, y) | (x, y) \in \Gamma_c \end{bmatrix}, \quad f(t) = \begin{bmatrix} f_1(t, x, y) | (x, y) \in \Gamma_c \\ f_2(t, x, y) | (x, y) \in \Gamma_c \end{bmatrix}.$$

Then the system (1) can be written as an evolution equation (compare [16]):

$$(5) \quad \frac{d^2 v}{dt^2} + Av(t) = Bf(t), \quad t \geq 0, \quad v(0) = v_0, \quad v_t(0) = v_1,$$

where

$$v_0 = \begin{bmatrix} u_0(x, y) \\ u_0 | \Gamma_c \\ (\partial u_0 / \partial n) | \Gamma_c \end{bmatrix} \quad \text{and} \quad v_1 = \begin{bmatrix} u_1 \\ u_1 | \Gamma_c \\ (\partial u_1 / \partial n) | \Gamma_c \end{bmatrix}.$$

Remark. Each classical solution of the initial-boundary-value problem (1) with smooth controllers f_1 and f_2 is clearly a solution of the evolution equation (5); conversely, each mild solution of (5) starting from smooth initial data yields a classical solution of (1). Hence for the approximate controllability and feedback stabilization of (1) it is sufficient to consider (5). Therefore, we shall interpret the control problems associated with classical solutions of (1) in terms of the corresponding control problems of the evolution equation (5) and, later on, of the first-order evolution equation (8).

LEMMA 1. *The operator $A: D(A) \rightarrow H$ is a densely defined, closable, symmetric, and coercively accretive operator. It admits a self-adjoint Friedrichs extension denoted by A_e , which has a compact resolvent.*

Proof (sketched). Through a straightforward calculation (see [17] for details) we can show that

$$\left\langle A \begin{bmatrix} u \\ u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} v \\ v_1 \\ v_2 \end{bmatrix} \right\rangle = \langle u, v \rangle = \left\langle \begin{bmatrix} u \\ u_1 \\ u_2 \end{bmatrix}, A \begin{bmatrix} v \\ v_1 \\ v_2 \end{bmatrix} \right\rangle,$$

where

$$\langle u, v \rangle = \iint_{\Omega} [\Delta u \Delta v + (1 - \sigma)(2u_{xy}v_{xy} - u_{xx}v_{yy} - u_{yy}v_{xx})] dx dy$$

for (u, u_1, u_2) and (v, v_1, v_2) in $D(A)$. It can be shown as in [17] that

$$\langle u, v \rangle = \langle A^{1/2}(u, u_1, u_2), A^{1/2}(v, v_1, v_2) \rangle.$$

Hence A is symmetric and coercively accretive. The symmetry and semibounded property imply that A admits a self-adjoint Friedrichs extension. Moreover, since

$$\|u_i\|_{L^2(\Gamma_c)} \leq \text{const} \|u\|_{H^2(\Omega)}, \quad i = 1, 2,$$

it can be shown by the Rellich theorem that both A and its extension A_e have a compact resolvent. \square

Denote $D(A^{1/2})$ with the graph norm by M , and define

$$X = M \times H,$$

which can be interpreted as the energy space. Now define two more operators:

$$(6) \quad G = \begin{bmatrix} 0 & I \\ -A_e & 0 \end{bmatrix} : D(G) \rightarrow X, \quad \text{where } D(G) = D(A) \times M,$$

$$(7) \quad K = \begin{bmatrix} 0 \\ B \end{bmatrix} : L^2(\Gamma_c) \times L^2(\Gamma_c) \rightarrow X.$$

The operator K is a bounded linear operator, and the operator G is the generator of a unitary group of linear operators $T(t)$ on X given by

$$T(t) = \begin{bmatrix} \cos(A^{1/2}t) & A^{-1/2} \sin(A^{1/2}t) \\ -A^{1/2} \sin(A^{1/2}t) & \cos(A^{1/2}t) \end{bmatrix}.$$

Let

$$w(t) = \begin{bmatrix} v(t) \\ (d/dt)v(t) \end{bmatrix} \quad \text{and} \quad w_0 = \begin{bmatrix} v_0 \\ v_1 \end{bmatrix},$$

where $(d/dt)v(t)$ stands for the strong derivative of $v(t)$ in H . Then the second-order evolution system (5) is reduced to the following first-order evolution system:

$$(8) \quad \begin{aligned} \frac{dw}{dt} &= Gw(t) + Kf(t), \quad t \geq 0, \\ w(0) &= w_0 \in X. \end{aligned}$$

The mild solution of the evolution equation (8) will be considered as the state function.

3. Energy decay with dissipative feedback. We introduce the feedback control

$$f = \begin{bmatrix} -(\partial u / \partial t)|_{\Gamma_c} \\ -(\partial u_n / \partial t)|_{\Gamma_c} \end{bmatrix}$$

as a dissipative perturbation of the generator G . The perturbed generator is denoted by

$$(9) \quad G_c = G - KK^*, \quad \text{with } D(G_c) = D(G).$$

Denote by $D(G^\infty)$ the dense subspace in X defined by

$$D(G^\infty) = \bigcap_{k=1}^{\infty} D(G^k),$$

where $D(G^k)$ is the domain of the operator G^k . Let the C_0 -contraction semigroup generated by G_c be $S(t)$, $t \geq 0$. As in [17, Lemmas 6 and 7], we can show that for each $w_0 \in D(G^\infty)$

- (i) $\|S(t)w_0\|$ and $\|G_c S(t)w_0\|$ are nonincreasing for $t \geq 0$;
- (ii) any trajectory $\{S(t)w_0; t \geq 0\}$ is located within a compact subset of X ;
- (iii) the ω -limit set associated with any initial data w_0 ,

$$\omega(w_0) = \bigcap_{\tau > 0} \left\{ \text{cl} \left[\bigcup_{t \geq \tau} S(t)w_0 \right] \right\},$$

is a nonempty compact subset of X ;

(iv) for each point $w_\infty \in \omega(w_0)$ the following limit holds:

$$\|w_\infty\| = \lim_{t \rightarrow \infty} \|S(t)w_0\| := E_\infty(w_0);$$

(v) for each point $w_\infty \in \omega(w_0)$ we have $w_\infty \in D(G^\infty)$ and

$$\|S(t)w_\infty\| = \|w_\infty\| = E_\infty(w_0) \quad \text{for } t \geq 0.$$

We shall show later in Lemma 5 that $E_\infty(w_0) = 0$.

LEMMA 2. *If $w_\infty \in \omega(w_0)$ and*

$$(10) \quad w_\infty(t) = S(t)w_\infty = \begin{bmatrix} v_\infty(t) \\ (d/dt)v_\infty(t) \end{bmatrix} = \begin{bmatrix} u_\infty(t, x, y) \\ u_\infty|_{\Gamma_c} \\ \partial u_\infty / \partial n|_{\Gamma_c} \\ \dot{u}_\infty(t, x, y) \\ \dot{u}_\infty|_{\Gamma_c} \\ \partial \dot{u}_\infty / \partial n|_{\Gamma_c} \end{bmatrix},$$

then $\varphi(t, x, y) := \dot{u}_\infty(t, x, y)$ is a classical solution of the following system:

$$(11) \quad \begin{aligned} \frac{\partial^2 \varphi}{\partial t^2} + \Delta^2 \varphi &= 0 && \text{in } R^+ \times \Omega, \\ \varphi = \frac{\partial \varphi}{\partial n} &= 0 && \text{on } R^+ \times \Gamma_0, \\ \Delta \varphi + (1 - \sigma)A_2 \varphi = \frac{\partial \Delta \varphi}{\partial n} + (1 - \sigma)A_1 \varphi &= 0 && \text{on } R^+ \times \Gamma_f, \\ \varphi = \frac{\partial \varphi}{\partial n} = \frac{\partial^2 \varphi}{\partial n^2} = \frac{\partial^3 \varphi}{\partial n^3} &= 0 && \text{on } R^+ \times \Gamma_c. \end{aligned}$$

Proof. By property (v) and the observation that $w_\infty(t) = S(t)w_\infty$ is a strong solution of the equation

$$(12) \quad \frac{dw}{dt} = (G - KK^*)w,$$

it follows that

$$\frac{d}{dt} \|w_\infty(t)\|^2 = -2\|K^*w_\infty(t)\|^2 = 0, \quad t \geq 0.$$

Hence

$$K^*w_\infty(t) = B^* \frac{d}{dt} v_\infty(t) = 0, \quad t \geq 0,$$

so that by (10)

$$(13) \quad \begin{aligned} \varphi(t, x, y)|_{\Gamma_c} &= 0, && t \geq 0, \\ \frac{\partial \varphi}{\partial n}(t, x, y)|_{\Gamma_c} &= 0, && t \geq 0. \end{aligned}$$

Classical regularity theory [13, § 6.6] shows that in this case $\varphi(t, x, y)$ is a sufficiently regular function of $(t, x, y) \in R^+ \times \Omega$ up to the boundary, that is, $\varphi \in C^\infty(\Omega)$ and has a C^∞ -extension in a neighborhood of each point on any smooth arc of Γ . Thus $\varphi(t, x, y)$ is a classical solution of the following system:

$$\begin{aligned}
 (14) \quad & \frac{\partial^2 \varphi}{\partial t^2} + \Delta^2 \varphi = 0 && \text{in } R^+ \times \Omega, \\
 & \varphi = \frac{\partial \varphi}{\partial n} = 0 && \text{on } R^+ \times \Gamma_0, \\
 & \Delta \varphi + (1 - \sigma) A_2 \varphi = \frac{\partial \Delta \varphi}{\partial n} + (1 - \sigma) A_1 \varphi = 0 && \text{on } R^+ \times \Gamma_f, \\
 & \varphi = \frac{\partial \varphi}{\partial n} = 0 && \text{on } R^+ \times \Gamma_c, \\
 & \Delta \varphi + (1 - \sigma) A_2 \varphi = \frac{\partial \Delta \varphi}{\partial n} + (1 - \sigma) A_1 \varphi = 0 && \text{on } R^+ \times \Gamma_c,
 \end{aligned}$$

according to (11) and (13). Since $\varphi = \partial \varphi / \partial n = 0$ along each arc of Γ_c for $t \geq 0$, we have

$$\frac{\partial \varphi}{\partial s} = \frac{\partial \varphi}{\partial n} = 0 \quad \text{on } R^+ \times \Gamma_c.$$

It is known [5] that on $R^+ \times \Gamma_c$

$$(15) \quad A_1 \varphi = \frac{\partial}{\partial s} \left[\frac{\partial^2 \varphi}{\partial n \partial s} - 2\kappa \frac{\partial \varphi}{\partial s} \right] = \frac{\partial^3 \varphi}{\partial s^2 \partial n} = 0,$$

and

$$(16) \quad A_2 \varphi = -\frac{\partial^2 \varphi}{\partial s^2} - \kappa \frac{\partial \varphi}{\partial n} = -\frac{\partial^2 \varphi}{\partial s^2} = 0$$

in terms of the curvature κ of Γ_c . Substituting (15) and (16) into the boundary conditions on $R^+ \times \Gamma_c$ in (14), we obtain

$$\varphi = \frac{\partial \varphi}{\partial n} = \Delta \varphi = \frac{\partial \Delta \varphi}{\partial n} = 0 \quad \text{on } R^+ \times \Gamma_c.$$

It follows that

$$(17) \quad \varphi = \frac{\partial \varphi}{\partial n} = \frac{\partial^2 \varphi}{\partial n^2} = \frac{\partial^3 \varphi}{\partial n^3} = 0 \quad \text{on } R^+ \times \Gamma_c.$$

Finally, (14) and (17) imply that $\varphi(t, x, y) = \dot{u}_\infty(t, x, y)$ is a classical solution of the system (11). \square

4. Global uniqueness. The boundary-value problem (11) for the two-dimensional Petrovsky equation is overdetermined. In order to assert that the only possible solution of (11) is the trivial solution $\varphi(t, x, y) = 0$ in $R^+ \times \Omega$, we shall invoke a global uniqueness theorem due to Hörmander.

LEMMA 3 [2]. *Let Q_1 and Q_2 be two open convex sets in R^n , such that $Q_2 \supset Q_1$. Let $P(D)$ be a differential operator with constant coefficients, such that every hyperplane that is characteristic with respect to $P(D)$ and intersects Q_2 also meets Q_1 . Then every*

distribution solution $u \in \mathcal{D}'(Q_2)$ satisfying $P(D)u = 0$ and vanishing in Q_1 must vanish in Q_2 .

Proof. See [2, Thm. 5.3.3].

LEMMA 4. The boundary-value problem (11) admits a unique solution $\varphi(t, x, y) = 0$ for $(t, x, y) \in R^+ \times \text{cl } \Omega$.

Proof. Let $P(D)$ be the Petrovsky operator: $P(D) = \partial^2 / \partial t^2 + \Delta^2$ in the 3-cylinder $Q = R^+ \times \Omega$. Since the associated characteristic vector is $\xi = (\xi_t, \xi_x, \xi_y) = (1, 0, 0)$, every characteristic 2-plane associated with $P(D)$ is parallel to the (x, y) plane.

Let $\varphi(t, x, y)$ be a solution of the boundary-value problem (11). Then $P(D)\varphi = 0$ and the last boundary condition in (11) implies that

$$(18) \quad \frac{\partial^k \varphi}{\partial n^k} = 0 \quad \text{on } \Sigma_c = R^+ \times \Gamma_c \quad \text{for all } k \geq 0.$$

Then the Holmgren local uniqueness theorem implies that $\varphi = 0$ identically in a neighborhood of Σ_c in Q . In fact, there is an open disk Ω_1 in Ω such that $\varphi = 0$ identically in $Q_1 = R^+ \times \Omega_1$.

Choose any point $p \in Q$ with its projection $\hat{p} \in \Omega$. Then there exists a finite chain $\{\Omega_j: j = 1, \dots, N = N(p)\}$ with the property that each Ω_j is an open disk in Ω centered at p_j with $R^2\text{-meas}(\Omega_j \cap \Omega_{j+1}) > 0$ for $j = 1, 2, \dots, N-1$ and with $p_N = \hat{p}$. Now use Lemma 3 to examine $\varphi(t, x, y)$ on the corresponding chain $Q_j = R^+ \times \Omega_j$ for $j = 1, 2, \dots, N$.

Consider the nonempty convex set $Q_1 \cap Q_2$ on which $\varphi = 0$ identically. Clearly, each characteristic 2-plane that meets Q_2 also meets $Q_1 \cap Q_2$, since these are each 3-cylinders in $Q = R^+ \times \Omega$. Thus we conclude by using Lemma 3, that $\varphi = 0$ identically in Q_2 . Continuing this argument along the chain $\{Q_j: j = 1, 2, \dots, N\}$, we finally conclude that $\varphi = 0$ identically in $Q_N = R^+ \times \Omega_N$, so $\varphi = 0$ at the point $p \in Q_N$. This completes the proof. \square

LEMMA 5. Let Ω be a plane region of finite genus as specified above. Then for each initial state $w_0 \in X$, it follows that the solution $w(t) = S(t)w_0$ of the evolution equation (8) satisfies

$$(19) \quad \lim_{t \rightarrow \infty} \|S(t)w_0\| = 0.$$

Proof. From Lemma 4 we have shown that

$$(20) \quad \varphi(t, x, y) = \frac{d}{dt} u_\infty(t, x, y) = 0 \quad \text{in } \text{cl } Q = \text{cl } (R^+ \times \Omega).$$

It follows that $u_\infty(t, x, y) = u_\infty(0, x, y)$ for all $t \geq 0$ and any $(x, y) \in \text{cl } \Omega$. Denote $u_\infty(0, x, y)$ by $\varphi_0(x, y)$. Then φ_0 satisfies the following boundary-value problem of the biharmonic equation:

$$(21) \quad \begin{aligned} \Delta^2 \varphi_0 &= 0 && \text{in } \Omega, \\ \varphi_0 = \frac{\partial \varphi_0}{\partial n} &= 0 && \text{on } \Gamma_0, \\ \frac{\partial \Delta \varphi_0}{\partial n} + (1 - \sigma)A_1 \varphi_0 &= 0 && \text{on } \Gamma_1, \\ \Delta \varphi_0 + (1 - \sigma)A_2 \varphi_0 &= 0 && \text{on } \Gamma_1. \end{aligned}$$

A straightforward calculation shows that

$$\begin{aligned} & \int \int_{\Omega} (\Delta^2 \varphi_0) \varphi_0 \, dx \, dy + \int_{\Gamma = \Gamma_0 \cup \Gamma_1} \left[(\Delta \varphi_0 + (1 - \sigma) A_2 \varphi_0) \frac{\partial \varphi_0}{\partial n} - \left(\frac{\partial \Delta \varphi_0}{\partial n} + (1 - \sigma) A_1 \varphi_0 \right) \varphi_0 \right] ds \\ &= \langle \varphi_0, \varphi_0 \rangle = \left\| A^{1/2} \begin{bmatrix} \varphi_0 \\ \varphi_0|_{\Gamma_c} \\ (\partial \varphi_0 / \partial n)|_{\Gamma_c} \end{bmatrix} \right\|^2 = 0. \end{aligned}$$

It follows that

$$\varphi_0(x, y) = 0 \quad \text{in } H^2(\Omega).$$

Therefore, we have the pointwise equality

$$u_{\infty}(t, x, y) = 0 \quad \text{in cl } Q,$$

so that

$$(22) \quad S(t)w_{\infty} = 0 \quad \text{for } t \geq 0.$$

Then by the properties of the ω -limit set $\omega(w_0)$, (22) implies that (19) holds. \square

5. Dynamic boundary stabilization. In this section we will achieve the main results in terms of stabilization and controllability for the evolution system (8) deduced from the original boundary control system (1) for an elastic plate of general shape.

THEOREM 1. *The evolution system (8), which models the elastic-plate system (1) over a region Ω of finite genus, is strongly stabilized by the following boundary-damping feedback:*

$$(23) \quad f_1(t, x, y) = -\frac{\partial u}{\partial t}(t, x, y) \quad \text{and} \quad f_2(t, x, y) = -\frac{\partial^2 u}{\partial t \partial n}(t, x, y),$$

for $t \geq 0$, $(x, y) \in \Gamma_c$.

Proof. From Lemma 5 it is known that the convergence (19) holds. This means that the linear feedback control

$$(24) \quad f(t) = -K^* w(t) = -B^* \frac{d}{dt} v(t) = - \begin{bmatrix} (\partial u / \partial t)|_{\Gamma_c} \\ (\partial^2 u / \partial t \partial n)|_{\Gamma_c} \end{bmatrix}, \quad t \geq 0,$$

strongly stabilizes the evolution system (8) because the solutions $S(t)w_0$ of the corresponding closed-loop system (12) strongly converge to the equilibrium at 0 for any initial data w_0 in the state space X . \square

THEOREM 2. *The evolution system (8), which models the elastic-plate system (1) over a region Ω of finite genus, is open-loop approximately controllable.*

Proof. By the controllability theory for linear systems in Hilbert spaces [1], [15], the strong stabilizability of the system (8) implies that it is approximately controllable in X by open-loop controllers f_1 and f_2 applied on Γ_c . Thus the conclusion of Theorem 2 is a consequence of Theorem 1. \square

Example. Let

$$\Omega = \{(r, \theta): \delta < r < 1, 0 \leq \theta < 2\pi\}, \quad 0 < \delta < 1 \text{ constant},$$

$$\Gamma_0 = \{(r, \theta): r = \delta\}, \quad \Gamma_1 = \{(r, \theta): r = 1\},$$

$$\Gamma_c = \{(r, \theta) \in \Gamma_1: \theta \in (\alpha, \beta) \text{ with } 0 < \beta - \alpha < 2\pi\},$$

$$\Gamma_f = \Gamma_1 \setminus \Gamma_c.$$

Consider a thin, isotropic, homogeneous, elastic annular plate over the region Ω . By the theory developed in this paper the corresponding plate system (1) is strongly stabilizable by the linear feedback (23) applied to the boundary arc Γ_c —no matter how small the controlled arc (α, β) is. As a consequence, this system is open-loop approximately controllable with the controllers supported on Γ_c .

REFERENCES

- [1] C. D. BENCHIMOL, *A note on weak stabilizability of contraction semigroups*, SIAM J. Control Optim., 16 (1978), pp. 373–379.
- [2] L. HÖRMANDER, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, New York, 1963.
- [3] J. E. LAGNESE, *Boundary Stabilization of Thin Plates*, Studies in Applied Mathematics, Vol. 10, SIAM Publications, Philadelphia, PA, 1989.
- [4] J. E. LAGNESE AND J. L. LIONS, *Modelling, Analysis and Control of Thin Plates*, Recherches en Mathématiques Appliquées, Vol. 6, Masson, Paris, 1988.
- [5] J. E. LAGNESE, *Recent progress in exact boundary controllability and uniform stabilizability of thin beams and plates*, in Distributed Parameter Control Systems—New Trends and Applications, G. Chen, E. B. Lee, W. Littman, and L. Markus, eds., Marcel Dekker, New York, 1991, pp. 61–111.
- [6] I. LASIECKA AND R. TRIGGIANI, *Exact controllability and uniform energy decay for Kirchhoff plates with boundary control only on $\Delta w|_{\Sigma}$* , J. Differential Equations, to appear.
- [7] —, *Exact controllability of the Euler–Bernoulli equation with controls in the Dirichlet and Neumann boundary conditions: a nonconservative case*, SIAM J. Control Optim., 27 (1989), pp. 330–373.
- [8] —, *Exact controllability of the Euler–Bernoulli equation with boundary controls for displacement and moment*, J. Math. Anal. Appl., 146 (1990), pp. 1–33.
- [9] J. L. LIONS, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, tome 1, Contrôlabilité exacte*, Recherches en Mathématiques Appliquées, Vol. 8, Masson, Paris, 1988.
- [10] W. LITTMAN AND L. MARKUS, *Exact boundary controllability of a hybrid system of elasticity*, Arch. Rational Mech. Anal., 103 (1988), pp. 193–236.
- [11] —, *Stabilization of a hybrid system of elasticity by feedback boundary damping*, Ann. Mat. Pura Appl., 152 (1988), pp. 281–330.
- [12] W. LITTMAN, L. MARKUS, AND Y. YOU, *A note on stabilization and controllability of a hybrid elastic system with boundary control*, Mathematics Report, University of Minnesota, Minneapolis, MN, 1987.
- [13] C. B. MORREY, JR., *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1966.
- [14] M. SCHIFFER AND S. BERGMAN, *Kernel Functions and Differential Equations*, Academic Press, New York, 1953.
- [15] M. SLEMROD, *A note on complete controllability and stabilizability for linear control systems in Hilbert spaces*, SIAM J. Control, 12 (1974), pp. 500–508.
- [16] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.
- [17] Y. YOU, *Boundary stabilization of two-dimensional Petrovsky equations: vibrating plate*, Differential Integral Equations, 4 (1991), pp. 617–638.
- [18] —, *Pointwise boundary stabilizability of hyperbolic evolution equations: two-dimensional hybrid elastic structures*, J. Math. Anal. Appl., 165 (1992), pp. 239–265.

BOUNDARY CONTROL OF SEMILINEAR ELLIPTIC EQUATIONS WITH POINTWISE STATE CONSTRAINTS*

EDUARDO CASAS[†]

Abstract. This paper is concerned with state constrained optimal control problems of semilinear elliptic equations, the control being on the boundary. Optimality conditions are derived and regularity of the optimal solution is investigated.

Key words. optimal control, boundary control, semilinear elliptic operators, optimality conditions, state constraints

AMS subject classifications. 49K20, 49J20

1. Introduction. This paper deals with state-constrained optimal control problems governed by a monotone semilinear and elliptic operator. A Neumann condition is considered, and the control should act on the boundary. Our aim is to derive the optimality conditions and to deduce some regularity results for the optimal control. This kind of optimal control problems arises in connection with some realistic problems; see Bermúdez and Martínez [2] and Luneville [12].

Optimal control problems with state constraints, governed by linear elliptic equations, have been studied by the author [6]–[9]. In these papers, the control was distributed in Ω or it was a coefficient of the operator. Abergel and Temam [1] studied some nonqualified problems of distributed control, deriving some nonstandard optimality conditions that become the usual ones when the problem is qualified. The semilinear case was considered by Bonnans and Casas [4], [5], the control being distributed in Ω , too. Optimality conditions for some boundary control problems with pointwise state constraints have been obtained by Mackenroth [13], [14] in the case of a linear operator with bounded controls; Luneville [12] in the case of Laplace operator and dimension 2 or 3 and the state subject to pointwise constraints only in a strict subset of Ω ; Bonnans and Casas [3] in the case of a particular semilinear equation. In the previous papers, except that of Luneville, the adjoint state equation is not investigated, and therefore regularity of the adjoint state was not deduced, which is essential to derive regularity results of optimal control.

In this paper, we show that, assuming the control to be in $L^t(\Gamma)$ for $t > n - 1$, $n > 1$ arbitrary, it is possible to derive the optimality conditions and to deduce some regularity results of the optimal control after having investigated regularity for the adjoint state.

The plan of this paper is as follows: in the next section, the control problem is formulated; in §3 the state equation is studied, in particular continuity of state is proved; in §4 we prove existence, uniqueness, and regularity of the solution of a Neumann problem governed by a linear elliptic equation with measures as data; finally, in §5 we apply this study to our control problem and derive the optimality conditions and regularity results.

2. Formulation of the control problem. Let Ω be an open bounded subset of \mathbb{R}^n ($n \geq 2$) with $C^{1,1}$ boundary Γ . Let us consider the following boundary value

* Received by the editors November 26, 1990; accepted for publication (in revised form) September 25, 1991. This research was supported in part by DGICYT (Madrid).

[†] Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. de Caminos, C. y P., Universidad de Cantabria, 39071-Santander, Spain.

problem:

$$(2.1) \quad \begin{aligned} Ay + \phi(y) &= f && \text{in } \Omega, \\ \partial_{\nu_A} y &= u && \text{on } \Gamma, \end{aligned}$$

with $f \in L^\rho(\Omega)$, $\rho > n/2$, $u \in L^t(\Gamma)$, $t > n - 1$, and

$$\begin{aligned} Ay &= - \sum_{i,j=1}^n \partial_{x_j} (a_{ij}(x) \partial_{x_i} y(x)) + a_0(x) y(x), \\ \partial_{\nu_A} y &= \sum_{i,j=1}^n a_{ij}(x) \partial_{x_i} y(x) \nu_j(x), \end{aligned}$$

where $\nu(x)$ denotes the unit outward normal to Γ at the point x ,

$$a_{ij} \in C^{0,1}(\overline{\Omega}) \quad \text{and} \quad a_0 \in L^\infty(\Omega),$$

$$(2.2) \quad \exists m > 0 \text{ such that } \sum_{i,j=1}^n a_{ij}(x) \xi_i \xi_j \geq m |\xi|^2 \quad \forall \xi \in R^n, x \in \Omega,$$

$$a_0(x) \geq 0 \quad \text{a.e. } x \in \Omega \text{ and } a_0 \not\equiv 0;$$

$$(2.3) \quad \begin{aligned} \phi : R &\longrightarrow R \text{ is of class } C^1, \\ \phi &\text{ is increasing monotone and } \phi(0) = 0. \end{aligned}$$

We consider the following control problem:

$$(P) \quad \begin{cases} \text{Minimize } J(u), \\ u \in K \text{ and } y_u \in C, \end{cases}$$

where K is a nonempty, convex, closed subset of $L^t(\Gamma)$; C is a closed and convex subset of $C(X)$ with nonempty interior, X being a compact subset of $\overline{\Omega}$; and the functional $J : L^t(\Gamma) \longrightarrow [0, +\infty]$ is defined by

$$J(u) = \frac{1}{2} \int_{\Omega} |y_u(x) - y_d(x)|^2 dx + \frac{N}{\sigma} \int_{\Gamma} |u(x)|^\sigma dm_{\Gamma}(x),$$

y_d being given in $L^2(\Omega)$, $\sigma \in (1, +\infty)$ and $N \geq 0$. Furthermore, we assume that one of the following hypotheses is satisfied:

(H1) K is bounded in $L^t(\Gamma)$, $t > n - 1$, and $\sigma \leq t$;

(H2) $N > 0$ and $\sigma \geq t > n - 1$.

As realistic examples, we can take

$$K = \{u \in L^\infty(\Gamma) : a \leq u(x) \leq b \text{ a.e. } [m_{\Gamma}] x \in \Gamma\}$$

and $\sigma = 2$, with $-\infty < a < b < +\infty$, which satisfies the first hypothesis, or

$$K = \{u \in L^t(\Gamma) : u(x) \geq 0 \text{ a.e. } [m_{\Gamma}] x \in \Gamma\}$$

and $\sigma = t$, which satisfies the second hypothesis. Above, m_{Γ} denotes the usual $(n - 1)$ -dimensional measure over Γ induced by the parametrization.

In the case where $n = 2$, we can take $\sigma = t = 2$, which is the most usual choice in control theory.

As examples of sets C , let us give

$$X = \overline{\Omega} \quad \text{and} \quad C = \{z \in C(\overline{\Omega}) : |z(x)| \leq \delta \quad \forall x \in \overline{\Omega}\},$$

$$X = \Gamma \quad \text{and} \quad C = \{z \in C(\Gamma) : |z(x)| \leq \delta \quad \forall x \in \Gamma\},$$

or

$$X \subset \Omega \quad \text{and} \quad C = \{z \in C(X) : |z(x)| \leq \delta \quad \forall x \in X\},$$

where $\delta > 0$ is a given constant.

Remark 1. In practice, the hypothesis $\phi(0) = 0$ is not a restriction because, if it is not verified, then it is enough to replace ϕ by $\phi - \phi(0)$ and f by $f - \phi(0)$, and so we have the required conditions.

3. Study of the state equation. In this section, we prove that the boundary problem (2.1) has a unique solution $y_u \in H^1(\Omega) \cap C(\overline{\Omega})$ for each $u \in L^t(\Gamma)$, $t > n - 1$, and that the relation between control and state is of class C^1 .

THEOREM 3.1. *For all $u \in L^t(\Gamma)$, with $t > n - 1$, and $f \in L^\rho(\Omega)$, $\rho > n/2$, there exists a unique solution y_u of the Neumann problem (2.1) belonging to $H^1(\Omega) \cap C(\overline{\Omega})$. Moreover, there exists a constant C_1 independent of f and u such that*

$$(3.1) \quad \|y_u\|_{H^1(\Omega)} + \|y_u\|_{C(\overline{\Omega})} \leq C_1 (\|f\|_{L^\rho(\Omega)} + \|u\|_{L^t(\Gamma)}).$$

Finally, if $\{u_k\} \subset L^t(\Gamma)$ converges weakly (or weak* in the case where $t = +\infty$) toward u in $L^t(\Gamma)$, then $\{y_{u_k}\}$ converges to y_u strongly in $H^1(\Omega) \cap C(\overline{\Omega})$.

Before proving this theorem, we must state the following lemma.

LEMMA 3.2. *Let us suppose that ϕ is a bounded function in R , $f \in L^\rho(\Omega)$ and $u \in W^{-1/r,r}(\Gamma)$, with $\rho > n/2$ and $r > n$. Then there exists a unique solution y_u in $H^1(\Omega) \cap L^\infty(\Omega)$ of the Neumann problem (2.1). Moreover, there exists a constant $C_2 > 0$ independent of u , f , and ϕ such that*

$$(3.2) \quad \|y_u\|_{H^1(\Omega)} + \|y_u\|_{L^\infty(\Omega)} \leq C_2 (\|f\|_{L^\rho(\Omega)} + \|u\|_{W^{-1/r,r}(\Gamma)}).$$

Proof. Utilizing the monotone operator theory (see Lions [11]), the existence and uniqueness of a solution $y_u \in H^1(\Omega)$ of (2.1) follows easily. On the other hand, arguing as in [17, Thm. 4.2], we obtain the boundedness of y_u and inequality (3.2). \square

Proof of Theorem 3.1. For every positive integer k , let us consider the functions $\phi_k : R \rightarrow R$ defined by

$$\phi_k(\theta) = \begin{cases} \phi(\theta) & \text{if } |\phi(\theta)| \leq k, \\ +k & \text{if } \phi(\theta) \geq +k, \\ -k & \text{if } \phi(\theta) \leq -k. \end{cases}$$

Then, due to Lemma 3.2, we know that the Neumann problem

$$\begin{aligned} Ay + \phi_k(y) &= f & \text{in } \Omega, \\ \partial_{\nu_A} y &= u & \text{on } \Gamma \end{aligned}$$

has a unique solution $y_k \in H^1(\Omega) \cap L^\infty(\Omega)$ and

$$(3.3) \quad \|y_k\|_{H^1(\Omega)} + \|y_k\|_{L^\infty(\Omega)} \leq C_2 (\|f\|_{L^\rho(\Omega)} + \|u\|_{W^{-1/r,r}(\Gamma)}),$$

where C_2 is independent of k . Therefore there exists a constant $c > 0$ such that $\|y_k\|_\infty \leq c$ for each k , which implies that there exists a positive integer $k_0 > 0$ such that $\phi_k(y_k) = \phi(y_k)$ for every $k \geq k_0$. That is, every y_k , with $k \geq k_0$, is a solution of the Neumann problem (2.1). On the other hand, the uniqueness of solution of this problem in $H^1(\Omega) \cap L^\infty(\Omega)$ is an immediate consequence of the properties of A and the monotonicity of ϕ .

Once we have proved existence, uniqueness, and the estimations in the norm $H^1(\Omega)$ and $L^\infty(\Omega)$ (consequence of inequality (3.3)), it remains to establish the continuity of y_u . For this, we utilize the density of $D(\Gamma)$ and $D(\Omega)$ in $L^t(\Gamma)$ and $L^\rho(\Omega)$, respectively. Let $\{u_k\} \subset D(\Gamma)$ and $\{f_k\} \subset D(\Omega)$ be two sequences converging to u and f in the topology of $L^t(\Gamma)$ and $L^\rho(\Omega)$, respectively, and let $y_k = y_{u_k}$ be the solution of the Neumann problem (2.1) corresponding to u_k and f_k . From the hypotheses on the regularity of A and Γ , it follows (see, for example, Grisvard [10]) that $\{y_k\} \subset W^{2,r}(\Omega) \subset C(\bar{\Omega})$. Now let us take $z_k = y_u - y_k$, $v_k = u - u_k$ and $g_k = f - f_k$, then z_k verifies that

$$\begin{aligned} Az_k + \alpha_k(x)z_k &= g_k & \text{in } \Omega, \\ \partial_{\nu_A} z_k &= v_k & \text{on } \Gamma, \end{aligned}$$

where

$$\alpha_k(x) = \begin{cases} \frac{\phi(y_u(x)) - \phi(y_k(x))}{y_u(x) - y_k(x)} & \text{if } y_u(x) \neq y_k(x), \\ 0 & \text{if } y_u(x) = y_k(x). \end{cases}$$

The monotonicity of ϕ implies the positivity of α_k . Applying Lemma 3.2, we obtain that

$$\|z_k\|_{L^\infty(\Omega)} \leq C_2 (\|g_k\|_{L^\rho(\Omega)} + \|v_k\|_{L^t(\Gamma)}) \rightarrow 0 \quad \text{if } k \rightarrow \infty,$$

which proves the uniform convergence of $\{y_k\}$ toward y_u and hence the continuity of y_u .

Finally, to prove the continuous dependence of y_u with respect to u , note that the weak convergence of $\{u_k\}$ to u in $L^t(\Gamma)$ implies the strong convergence in $W^{-1/r,r}(\Gamma)$. Therefore it is enough to again take $z_k = y_u - y_{u_k}$ and to argue as above to conclude the strong convergence of $\{y_{u_k}\}$ to y_u in $H^1(\Omega) \cap C(\bar{\Omega})$. \square

We finish this section proving that the relation between the control and the state is differentiable.

THEOREM 3.3. *The mapping $G : L^t(\Gamma) \rightarrow H^1(\Omega) \cap C(\bar{\Omega})$ defined by $G(u) = y_u$ is of class C^1 , and, for every $u, v \in L^t(\Gamma)$, the element $z = DG(u) \cdot v$ is the unique solution of the Neumann problem*

$$(3.4) \quad \begin{aligned} Az + \phi'(y_u)z &= 0 & \text{in } \Omega, \\ \partial_{\nu_A} z &= v & \text{on } \Gamma. \end{aligned}$$

Proof. First, we prove that G is Gâteaux differentiable. For it, let us take for every $h > 0$, y_h as the solution of the Neumann problem

$$\begin{aligned} Ay + \phi(y) &= f & \text{in } \Omega, \\ \partial_{\nu_A} y &= u + hv & \text{on } \Gamma. \end{aligned}$$

Then $z_h = (y_h - y_u)/h$ satisfies

$$\begin{aligned} Az_h + \alpha_h(x)z_h &= 0 & \text{in } \Omega, \\ \partial_{\nu_A} z_h &= v & \text{on } \Gamma, \end{aligned}$$

where $\alpha_h(x) = \phi'(y_u(x) + \theta_h(x)[y_h(x) - y_u(x)])$, with $\theta_h(x) \in (0, 1)$. Using Theorem 3.1, we obtain that $\{z_h\}$ is uniformly bounded in $H^1(\Omega) \cap C(\overline{\Omega})$. Therefore it is easy to pass to the limit in the previous problem and to deduce the convergence of $\{z_h\}$ to z in $H^1(\Omega) \cap C(\overline{\Omega})$, where z is the solution of problem (3.4). On the other hand, it is a consequence of Theorem 3.1 that the linear mapping $v \rightarrow z$ is continuous from $L^t(\Gamma)$ to $H^1(\Omega) \cap C(\overline{\Omega})$. Finally, due to the continuity of ϕ' , it follows easily that DG is continuous, which concludes the proof. \square

4. Study of a Neumann problem with measures as data. As we see in the next section, the adjoint state equation of the optimality system for our control problem has measures as data in Ω and on Γ . In this section, we study this type of boundary value problems, proving existence and uniqueness of a solution, deriving a regularity result, and stating a trace theorem and a Green's formula.

The model problem is the following:

$$(4.1) \quad \begin{aligned} Ap &= \mu_\Omega & \text{in } \Omega, \\ \partial_{\nu_A} p &= \mu_\Gamma & \text{on } \Gamma, \end{aligned}$$

where A is an elliptic operator as defined in §2, satisfying the conditions (2.2), and μ_Ω and μ_Γ are real regular Borel measures in Ω and Γ , respectively, Ω and Γ verifying also the conditions stated in §2.

We begin by establishing a trace theorem, but first it is necessary to introduce some function spaces. For every $s \in (1, n/(n-1))$, let us consider the space

$$V^s(\Omega) = \{\vec{p} \in L^s(\Omega)^n : \operatorname{div} \vec{p} \in M(\Omega)\},$$

where $M(\Omega)$ is the space formed by the real regular Borel measures in Ω . Endowed with the norm

$$\|\vec{p}\|_{V^s(\Omega)} = \|\vec{p}\|_{L^s(\Omega)^n} + \|\operatorname{div} \vec{p}\|_{M(\Omega)},$$

$V^s(\Omega)$ is a Banach space.

We will henceforth follow the notation

$$\langle \mu_\Omega, y \rangle_\Omega = \int_\Omega y(x) d\mu_\Omega(x) \quad \text{and} \quad \langle \mu_\Gamma, z \rangle_\Gamma = \int_\Gamma z(x) d\mu_\Gamma$$

for all functions $y \in C(\overline{\Omega})$ and $z \in C(\Gamma)$ and all real regular Borel measures $\mu_\Omega \in M(\Omega)$ and $\mu_\Gamma \in M(\Gamma)$.

We now have the following result.

THEOREM 4.1. *There exists a unique linear and continuous mapping*

$$\gamma_\nu : V^s(\Omega) \longrightarrow W^{-1/s, s}(\Gamma)$$

verifying

$$(4.2) \quad \gamma_\nu(\vec{p}) = \vec{p}|_\Gamma \cdot \vec{\nu} \quad \forall \vec{p} \in C^1(\overline{\Omega})^n,$$

$$(4.3) \quad \langle \gamma_\nu(\vec{p}), \gamma(z) \rangle = \int_\Omega \vec{p} \cdot \nabla z dx + \langle \operatorname{div} \vec{p}, z \rangle_\Omega \quad \forall z \in W^{1,r}(\Omega),$$

where r is the conjugate of s .

Proof. Let us take $g \in W^{1/s,r}(\Gamma) = \gamma(W^{1,r}(\Omega))$ and $z \in W^{1,r}(\Omega)$ such that $\gamma(z) = g$. Then we define

$$\langle \gamma_\nu(\vec{p}), g \rangle = \int_\Omega \vec{p} \cdot \nabla z dx + \langle \operatorname{div} \vec{p}, z \rangle_\Omega.$$

Let us prove that γ_ν is well defined. First, from the inequality $s < n/(n-1)$, it follows that $r > n$, and therefore $W^{1,r}(\Omega) \subset C(\overline{\Omega})$. On the other hand, if $z_1, z_2 \in W^{1,r}(\Omega)$ and $\gamma(z_1) = \gamma(z_2) = g$, then we must prove that

$$\int_\Omega \vec{p} \cdot \nabla z_1 dx + \langle \operatorname{div} \vec{p}, z_1 \rangle_\Omega = \int_\Omega \vec{p} \cdot \nabla z_2 dx + \langle \operatorname{div} \vec{p}, z_2 \rangle_\Omega.$$

To do this, let us take $z = z_1 - z_2 \in W_0^{1,r}(\Omega)$ and $\{z_k\} \subset D(\Omega)$ a sequence converging to z in $W_0^{1,r}(\Omega)$. Since $r > n$, we have that $\nabla z_k \rightarrow \nabla z$ in $L^r(\Omega)^n$ and $z_k \rightarrow z$ in $C(\overline{\Omega})$, from where we obtain that

$$\int_\Omega \vec{p} \cdot \nabla z dx + \langle \operatorname{div} \vec{p}, z \rangle_\Omega = \lim_{k \rightarrow \infty} \left\{ \int_\Omega \vec{p} \cdot \nabla z_k dx + \langle \operatorname{div} \vec{p}, z_k \rangle_{M(\Omega), C(\Omega)} \right\} = 0,$$

the last equality being a consequence of the definition of derivation in the distribution sense.

So we have that γ_ν is well defined, and, obviously, it is linear. Let us prove the continuity

$$|\langle \gamma_\nu(\vec{p}), g \rangle| \leq \|\vec{p}\|_{L^s(\Omega)^n} \|\nabla z\|_{L^r(\Omega)^n} + \|\operatorname{div} \vec{p}\|_{M(\Omega)} \|z\|_{C(\overline{\Omega})} \leq c \|\vec{p}\|_{V^s(\Omega)} \|z\|_{W^{1,r}(\Omega)}.$$

Taking now the infimum we obtain that

$$|\langle \gamma_\nu(\vec{p}), g \rangle| \leq c \|\vec{p}\|_{V^s(\Omega)} \inf_{\gamma(z)=g} \|z\|_{W^{1,r}(\Omega)} = c \|\vec{p}\|_{V^s(\Omega)} \|g\|_{W^{1/s,r}(\Gamma)},$$

which implies the continuity of γ_ν .

From the definition of γ_ν and using the Green's formula for regular functions, it is immediate to prove that (4.2) is satisfied. The uniqueness follows from (4.3) and the surjectivity of $\gamma : W^{1,r}(\Omega) \rightarrow W^{1/s,r}(\Gamma)$. \square

DEFINITION 4.2. Given $p \in W^{1,s}(\Omega)$, satisfying that $Ap \in M(\Omega)$, we define $\partial_{\nu_A} p = \gamma_\nu(\vec{w})$, where \vec{w} is given by

$$(4.4) \quad w_j(x) = \sum_{i=1}^n a_{ij}(x) \partial_{x_i} p(x), \quad 1 \leq j \leq n.$$

Let us note that $\vec{w} \in L^s(\Omega)^n$ and

$$\operatorname{div} \vec{w} = -Ap + a_0(x)p \in M(\Omega),$$

which implies that $\vec{w} \in V^s(\Omega)$. Hence $\gamma_\nu(\vec{w})$ is well defined as an element of the space $W^{-1/s,s}(\Gamma)$. From Theorem 4.1, we deduce that the previous definition agrees with

the usual definition when p is a regular function. We now study the existence and uniqueness of a solution of problem (4.1) in the space $W^{1,s}(\Omega)$.

THEOREM 4.3. *The Neumann problem (4.1) has a unique solution belonging to the space $W^{1,s}(\Omega)$ for every $s \in [1, n/(n-1))$. Furthermore, the following inequality is verified:*

$$(4.5) \quad \|p\|_{W^{1,s}(\Omega)} \leq C_3(\|\mu_\Omega\|_{M(\Omega)} + \|\mu_\Gamma\|_{M(\Gamma)}),$$

for some positive constant C_3 depending only on A and Ω .

Proof. Let $L : W^{2,r}(\Omega) \longrightarrow L^r(\Omega) \times W^{1/s,r}(\Gamma)$ be the operator defined by $L[z] = (A^*z, \partial_{\nu_{A^*}}z)$, where A^* is the adjoint operator of A . Due to the hypotheses on A and Γ , we have that L is an isomorphism; see Grisvard [10]. Then the adjoint operator $L^* : L^s(\Omega) \times W^{-1/s,s}(\Gamma) \longrightarrow (W^{2,r}(\Omega))'$ is also an isomorphism. Therefore, given $\mu = \mu_\Omega + \mu_\Gamma \in (W^{2,r}(\Omega))'$,

$$\langle \mu, z \rangle = \int_{\Omega} z(x) d\mu_\Omega(x) + \int_{\Gamma} z(x) d\mu_\Gamma(x),$$

there exists a unique element $(p, q) \in L^s(\Omega) \times W^{-1/s,s}(\Gamma)$ such that $L^*[(p, q)] = \mu$; that is,

$$(4.6) \quad \int_{\Omega} p A^* z dx + \langle q, \partial_{\nu_{A^*}} z \rangle = \int_{\Omega} z(x) d\mu_\Omega(x) + \int_{\Gamma} z(x) d\mu_\Gamma(x)$$

for all $z \in W^{2,r}(\Omega)$.

If we take $z \in D(\Omega)$ in the previous equation, we obtain that

$$(4.7) \quad Ap = \mu_\Omega \quad \text{in } \Omega.$$

Let us prove that $p \in W^{1,s}(\Omega)$. For every $\psi \in D(\Omega)$, let $z \in W^{2,r}(\Omega)$ be the solution of the Neumann problem

$$\begin{aligned} A^*z &= \partial_{x_j}\psi & \text{in } \Omega, \\ \partial_{\nu_{A^*}}z &= 0 & \text{on } \Gamma, \end{aligned}$$

with $1 \leq j \leq n$. Then, by using again [17, Thm. 4.2], we obtain that

$$\|z\|_{C(\bar{\Omega})} \leq c_1 \|\partial_{x_j}\psi\|_{W^{-1,r}(\Omega)} \leq c_1 \|\psi\|_{L^r(\Omega)}.$$

Using this inequality and (4.6), we obtain that

$$\begin{aligned} |\langle \partial_{x_j}p, \psi \rangle| &= \left| \int_{\Omega} p \partial_{x_j}\psi dx \right| = \left| \int_{\Omega} p A^*z dx \right| = \left| \int_{\Omega} z(x) d\mu_\Omega(x) + \int_{\Gamma} z(x) d\mu_\Gamma(x) \right| \\ &\leq (\|\mu_\Omega\|_{M(\Omega)} + \|\mu_\Gamma\|_{M(\Gamma)}) \|z\|_{C(\bar{\Omega})} \leq c_2 \|\psi\|_{L^r(\Omega)}, \end{aligned}$$

from where it follows, due to the density of $D(\Omega)$ in $L^r(\Omega)$, that $\partial_{x_j}p \in L^s(\Omega)$, $1 \leq j \leq n$. Hence we have that $p \in W^{1,s}(\Omega)$, and (4.5) follows from the previous inequality and the continuity of L^{-1} . Then, due to (4.7), $\partial_{\nu_{A^*}}p$ is well defined as an element of $W^{-1/s,s}(\Gamma)$; see Definition 4.2. Moreover, from (4.3), it follows that

$$(4.8) \quad \langle \partial_{\nu_{A^*}}p, \gamma(z) \rangle = a(p, z) - \int_{\Omega} z d\mu_\Omega \quad \forall z \in W^{1,r}(\Omega).$$

Now let us see that $\partial_{\nu_A} p = \mu_\Gamma$. First, we note that $\gamma(W^{1,r}(\Omega)) \subset C(\Gamma)$, and hence $M(\Gamma) \subset W^{-1/s,s}(\Gamma)$. Let $g \in W^{1+1/s,r}(\Gamma)$ arbitrary, and let us take $z \in W^{2,r}(\Omega)$ such that $\gamma(z) = g$ and $\partial_{\nu_{A^*}} z = 0$. This is possible because the mapping

$$\begin{array}{ccc} W^{2,r}(\Omega) & \longrightarrow & W^{1+1/s,r}(\Gamma) \times W^{1/s,r}(\Gamma), \\ z & \longrightarrow & (\gamma(z), \partial_{\nu_{A^*}} z) \end{array}$$

is surjective; see Lemma 4.4, proved below. From (4.8), integrating by parts and using (4.6), we obtain that

$$\begin{aligned} \langle \partial_{\nu_A} p, g \rangle &= \langle \partial_{\nu_A} p, \gamma(z) \rangle = a(p, z) - \int_{\Omega} z d\mu_{\Omega} \\ &= \int_{\Omega} p A^* z dx - \int_{\Omega} z d\mu_{\Omega} = \int_{\Gamma} z d\mu_{\Gamma} = \int_{\Gamma} g d\mu_{\Gamma}, \end{aligned}$$

which allows us to conclude that $\partial_{\nu_A} p$ and μ_Γ coincide over $W^{1+1/s,r}(\Gamma)$. Finally, from the density of $W^{1+1/s,r}(\Gamma)$ in $W^{1/s,r}(\Gamma)$, it follows the searched equality $\partial_{\nu_A} p = \mu_\Gamma$. Thus $p \in W^{1,s}(\Omega)$ is a solution of the Neumann problem (4.1). To finish this proof, it remains to state the unicity of solution.

Taking $g \in W^{1/s,r}(\Gamma)$ arbitrary and $z \in W^{2,r}(\Omega)$ such that $\gamma(z) = 0$ and $\partial_{\nu_{A^*}} z = g$, again using (4.8), (4.6), and integrating by parts, we obtain that

$$\begin{aligned} \int_{\Omega} z d\mu_{\Omega} &= a(p, z) = \int_{\Omega} p A^* z dx + \int_{\Gamma} \partial_{\nu_{A^*}} z \gamma(p) dm_{\Gamma}(x) \\ &= \int_{\Omega} z d\mu_{\Omega} - \langle q, \partial_{\nu_{A^*}} z \rangle + \int_{\Gamma} \partial_{\nu_{A^*}} z \gamma(p) dm_{\Gamma}(x); \end{aligned}$$

hence

$$\langle q, g \rangle = \int_{\Gamma} g \gamma(p) dm_{\Gamma}(x) \quad \forall g \in W^{1/s,r}(\Gamma),$$

which implies the equality $\gamma(p) = q$.

The uniqueness of solution of (4.1) in $W^{1,s}(\Omega)$ is obtained because every solution of (4.1) satisfies (4.6), which follows easily by using (4.3) and integrating by parts. \square

LEMMA 4.4. *The mapping*

$$\begin{array}{ccc} W^{2,r}(\Omega) & \longrightarrow & W^{1+1/s,r}(\Gamma) \times W^{1/s,r}(\Gamma), \\ z & \longrightarrow & (\gamma(z), \partial_{\nu_{A^*}} z) \end{array}$$

is surjective, where

$$\partial_{\nu_{A^*}} z = \sum_{i,j=1}^n a_{ji}(x) \partial_{x_j} z(x) \nu_i(x).$$

Proof. Let $(g, h) \in W^{1+1/s,r}(\Gamma) \times W^{1/s,r}(\Gamma)$. The vector $\nu_{A^*}(x)$ can be written in the form

$$\nu_{A^*}(x) = \alpha(x) \nu(x) + t(x),$$

where $t(x)$ is a tangent vector to Γ at the point x and

$$\alpha(x) = \nu_{A^*}(x) \cdot \nu(x) = \nu(x)^T (a_{ij}(x)) \nu(x) \geq m > 0.$$

Since Γ is of class $C^{1,1}$, $\nu(x)$ is a Lipschitz function on Γ . Therefore, remembering that the coefficients a_{ij} are Lipschitz, we deduce that α is a strictly positive Lipschitz function on Γ . Furthermore, $t(x)$ is also a Lipschitz function on Γ . Hence $(h - \partial_t g)/\alpha$ belongs to $W^{1/s,r}(\Gamma)$.

On the other hand, it is known that the mapping

$$\begin{array}{ccc} W^{2,r}(\Omega) & \longrightarrow & W^{1+1/s,r}(\Gamma) \times W^{1/s,r}(\Gamma), \\ z & \longrightarrow & (\gamma(z), \partial_\nu z) \end{array}$$

is surjective; see Grisvard [10] and Nečas [15]. Therefore there exists an element $z \in W^{2,r}(\Omega)$ such that $\gamma(z) = g$ and $\partial_\nu z = (h - \partial_t g)/\alpha$. Then we have that

$$\partial_{\nu_{A^*}} z(x) = \alpha(x) \partial_\nu z(x) + \partial_t z(x) = \alpha(x) \left[\frac{1}{\alpha(x)} (h(x) - \partial_t g(x)) \right] + \partial_t g(x) = h(x);$$

thus z is the searched element. \square

5. Study of the control problem. The first point to be considered in the study of the control problem (P) is the existence of a solution, which is proved in the next theorem.

THEOREM 5.1. *Under the hypotheses assumed in §2 and supposing the existence of a feasible control (i.e., a control $u \in K$ such that $y_u \in C$), then problem (P) has at least one solution. Moreover, if ϕ is linear, then the solution is unique.*

Proof. Let $\{u_k\}$ be a minimizing sequence in K . Because of hypothesis (H1) or (H2), this sequence is bounded in $L^t(\Gamma)$ (respectively, $L^\sigma(\Gamma)$). Then we can extract a subsequence, denoted in the same way, converging weakly to an element $\bar{u} \in K$ (since K is convex and closed). Now, using Theorem 3.1, we deduce that $y_{u_k} \rightarrow y_{\bar{u}}$ in $H^1(\Omega) \cap C(\bar{\Omega})$. Since $y_{u_k} \in C$ for every k and C is closed in $C(X)$, $X \subset \bar{\Omega}$, the uniform convergence implies that $y_{\bar{u}} \in C$; therefore \bar{u} is also a feasible control. Finally, it is immediate to verify that $\liminf J(u_k) \geq J(\bar{u})$, which proves that \bar{u} is a solution of (P).

If ϕ is linear, then problem (P) is strictly convex, and therefore the solution must be unique. \square

To derive the optimality conditions for problem (P), we must establish a preliminary result, which constitutes an abstract theorem of existence of Lagrange multiplier. Before stating this theorem, let us note that, in this paper, according to the most usual convention, the Gâteaux differential of a mapping is assumed linear and continuous.

THEOREM 5.2. *Let U and Z be two Banach spaces, and $K \subset U$ and $C \subset Z$ two convex subsets, C having a nonempty interior. Let $\bar{u} \in K$ be a solution of the optimization problem*

$$(Q) \begin{cases} \text{Min } J(u), \\ u \in K \text{ and } G(u) \in C, \end{cases}$$

where $J : U \rightarrow (-\infty, +\infty]$ and $G : U \rightarrow Z$ are two Gâteaux differentiable mappings at \bar{u} . Then there exist a real number $\bar{\lambda} \geq 0$ and an element $\bar{\mu} \in Z'$ such that

$$(5.1) \quad \bar{\lambda} + \|\bar{\mu}\|_{Z'} > 0,$$

$$(5.2) \quad \langle \bar{\mu}, z - G(\bar{u}) \rangle \leq 0 \quad \forall z \in C,$$

$$(5.3) \quad \langle \bar{\lambda} J'(\bar{u}) + [DG(\bar{u})]^* \bar{\mu}, u - \bar{u} \rangle \geq 0 \quad \forall u \in K.$$

Moreover, $\bar{\lambda}$ can be taken equal to 1 if the following condition of Slater type is satisfied:

$$(5.4) \quad \exists u_0 \in K \text{ such that } G(\bar{u}) + DG(\bar{u}) \cdot (u_0 - \bar{u}) \in \overset{\circ}{C}.$$

Proof. Let us consider the sets

$$A = \{(z, \lambda) \in Z \times R : \exists u \in K / z = G(\bar{u}) + DG(\bar{u}) \cdot (u - \bar{u}), \lambda = J'(\bar{u}) \cdot (u - \bar{u})\}$$

and $B = \overset{\circ}{C} \times (-\infty, 0)$. Using the linearity of $J'(\bar{u})$ and $DG(\bar{u})$, it is immediate to verify that A and B are convex sets. Moreover, they are disjoint. To see this, suppose that it is false; then there exists a point $u_0 \in K$ such that

$$\begin{aligned} z_0 &= G(\bar{u}) + DG(\bar{u}) \cdot (u_0 - \bar{u}) = G(\bar{u}) + \lim_{h \rightarrow 0} \frac{1}{h} [G(\bar{u} + h(u_0 - \bar{u})) - G(\bar{u})] \in \overset{\circ}{C}, \\ \lambda_0 &= J'(\bar{u}) \cdot (u_0 - \bar{u}) = \lim_{h \rightarrow 0} \frac{1}{h} (J(\bar{u} + h(u_0 - \bar{u})) - J(\bar{u})) < 0. \end{aligned}$$

Hence we can get a number $h_0 \in (0, 1)$ such that

$$\begin{aligned} z_h &= G(\bar{u}) + \frac{1}{h} (G(\bar{u} + h(u_0 - \bar{u})) - G(\bar{u})) \in \overset{\circ}{C} \quad \forall h \in (0, h_0), \\ \frac{1}{h} (J(\bar{u} + h(u_0 - \bar{u})) - J(\bar{u})) &< 0 \quad \forall h \in (0, h_0). \end{aligned}$$

Then we deduce that

$$G(\bar{u} + h(u_0 - \bar{u})) = h z_h + (1 - h) G(\bar{u}) \in \overset{\circ}{C} \quad \text{and} \quad J(\bar{u} + h(u_0 - \bar{u})) < J(\bar{u})$$

for every $h \in (0, h_0)$, which contradicts the fact that \bar{u} is a solution of (Q).

Now considering that B is an open set, from the geometric version of Hahn-Banach theorem, we deduce the existence of $\bar{\mu} \in Z'$ and $\bar{\lambda} \in R$, verifying that

$$(5.5) \quad \langle \bar{\mu}, z_1 \rangle + \bar{\lambda} \lambda_1 > \langle \bar{\mu}, z_2 \rangle + \bar{\lambda} \lambda_2 \quad \forall (z_1, \lambda_1) \in A, (z_2, \lambda_2) \in B.$$

Let us prove that $\bar{\lambda} \geq 0$. If $\bar{\lambda}$ were strictly negative, taking $\lambda_1 = 0$, $z_1 = G(\bar{u})$, $z_2 \in \overset{\circ}{C}$ fixed, and $\lambda_2 = -k$ in (5.5), with k positive integer, it follows that

$$\langle \bar{\mu}, G(\bar{u}) \rangle > \langle \bar{\mu}, z_2 \rangle - \bar{\lambda} k.$$

When k tends to infinity we arrive at a contradiction. Therefore $\bar{\lambda} \geq 0$. Furthermore, since inequality (5.5) is strict, $\bar{\lambda} = \bar{\mu} = 0$ is not an admissible possibility, which proves (5.1).

Now, from (5.5), since $\bar{B} = \bar{C} \times (-\infty, 0]$, we obtain that

$$(5.6) \quad \langle \bar{\mu}, z_1 \rangle + \bar{\lambda} \lambda_1 \geq \langle \bar{\mu}, z_2 \rangle + \bar{\lambda} \lambda_2 \quad \forall (z_1, \lambda_1) \in A, (z_2, \lambda_2) \in \bar{B}.$$

Thus it is enough to take $z_1 = G(\bar{u})$, $z_2 = z \in C$ and $\lambda_1 = \lambda_2 = 0$ to deduce (5.2). Inequality (5.3) is obtained by taking $z_1 = G(\bar{u}) + DG(\bar{u}) \cdot (u - \bar{u})$, $\lambda_1 = J'(\bar{u}) \cdot (u - \bar{u})$, with $u \in K$, $\lambda_2 = 0$ and $z_2 = G(\bar{u})$.

Finally, let us prove that $\bar{\lambda} \neq 0$ when the Slater condition is satisfied. Assume that (5.4) is verified and $\bar{\lambda} = 0$. As a first consequence, we have that

$$(5.7) \quad \langle \bar{\mu}, z - G(\bar{u}) \rangle < 0 \quad \forall z \in \overset{\circ}{C}.$$

To prove this inequality, it is enough to suppose that there exists $z_0 \in \overset{\circ}{C}$ such that $\langle \bar{\mu}, z_0 - G(\bar{u}) \rangle = 0$. Due to (5.2), we obtain that

$$\langle \bar{\mu}, z + z_0 - G(\bar{u}) \rangle \leq 0 \quad \forall z \in B_\epsilon(0),$$

with $\epsilon > 0$ small enough in a such way that $B_\epsilon(z_0) \subset \overset{\circ}{C}$. Hence $\langle \bar{\mu}, z \rangle \leq 0$ for every $z \in B_\epsilon(z_0)$, which implies that $\bar{\mu} = 0$, contradicting (5.1).

Now, taking $z = G(\bar{u}) + DG(\bar{u}) \cdot (u_0 - \bar{u}) \in \overset{\circ}{C}$ in (5.7), it follows that

$$\langle [DG(\bar{u})]^* \bar{\mu}, u_0 - \bar{u} \rangle = \langle \bar{\mu}, DG(\bar{u}) \cdot (u_0 - \bar{u}) \rangle < 0,$$

which contradicts (5.3); therefore $\bar{\lambda} > 0$. It is enough to divide (5.2) and (5.3) by $\bar{\lambda}$ and to again denote the quotient $\bar{\mu}/\bar{\lambda}$ by $\bar{\mu}$ to get the desired result. \square

Remark 2. If U is separable and J and G are of class C^1 in a neighborhood of \bar{u} , it is possible to prove, using the Clarke's generalized gradient calculus, that problem (Q) is normal almost always; this means that almost always we can take $\bar{\lambda} = 1$. More precisely, for all $\delta > 0$, let $C_\delta = (1 - \delta)z_0 + \delta C$ and

$$(Q_\delta) \begin{cases} \text{Min } J(u), \\ u \in K \text{ and } G(u) \in C_\delta. \end{cases}$$

If this problem has a solution for all number δ ranging an interval I , then (Q_δ) is normal for almost every $\delta \in I$; see Bonnans and Casas [3], [4].

Before applying the previous theorem to our control problem, let us introduce some notation. $M(X)$ denotes the space of real regular Borel measures in X ; that is, $M(X)$ is the dual space of $C(X)$ (note that X is a compact subset of $\bar{\Omega}$). The norm in $M(X)$ is given by

$$(5.8) \quad \|\mu\|_{M(X)} = |\mu|(X) = \sup \left\{ \int_X y(x) d\mu(x) : y \in C(X) \text{ and } \|y\|_\infty \leq 1 \right\},$$

where $|\mu|$ is the total variation measure; see Rudin [16]. Every element $\mu \in M(X)$ can be decomposed as a sum of two measures $\mu = \mu_\Omega + \mu_\Gamma$, μ_Ω and μ_Γ being regular real Borel measures in $\bar{\Omega}$, concentrated in $\Omega \cap X$ and $\Gamma \cap X$, respectively.

We now are ready to derive the optimality system for problem (P).

THEOREM 5.3. *Let \bar{u} be a solution of problem (P); then there exist a real number $\bar{\lambda} \geq 0$ and elements $\bar{y} \in H^1(\Omega) \cap C(\bar{\Omega})$, $\bar{p} \in W^{1,s}(\Omega)$ for all $s < n/(n-1)$ and $\bar{\mu} \in M(X)$ satisfying*

$$(5.9) \quad \bar{\lambda} + \|\bar{\mu}\|_{M(X)} > 0;$$

$$(5.10) \quad \begin{aligned} A\bar{y} + \phi(\bar{y}) &= f && \text{in } \Omega, \\ \partial_{\nu_A} \bar{y} &= \bar{u} && \text{on } \Gamma; \end{aligned}$$

$$(5.11) \quad \begin{aligned} A^* \bar{p} + \phi'(\bar{y}) \bar{p} &= \bar{\lambda}(\bar{y} - y_d) + \bar{\mu}_\Omega && \text{in } \Omega, \\ \partial_{\nu_A^*} \bar{p} &= \bar{\mu}_\Gamma && \text{on } \Gamma; \end{aligned}$$

$$(5.12) \quad \int_X (z - \bar{y}) d\bar{\mu} \leq 0 \quad \forall z \in C;$$

$$(5.13) \quad \int_{\Gamma} (\bar{p} + \bar{\lambda} N |\bar{u}|^{\sigma-2} \bar{u}) (u - \bar{u}) dm_{\Gamma}(x) \geq 0 \quad \forall u \in K,$$

where A^* is the adjoint operator of A . Moreover, if the following Slater condition is verified:

$$(5.14) \quad \exists (u_0, z_0) \in K \times (H^1(\Omega) \cap C(\bar{\Omega})) / (\bar{y} + z_0) \in \overset{\circ}{C},$$

where z_0 is the solution of the following Neumann problem:

$$\begin{aligned} Az_0 + \phi'(\bar{y})z_0 &= 0 & \text{in } \Omega, \\ \partial_{\nu_A} z_0 &= u_0 - \bar{u} & \text{on } \Gamma, \end{aligned}$$

then system (5.10)–(5.13) is satisfied with $\bar{\lambda} = 1$.

Proof. Theorem 5.3 is a consequence of Theorem 5.2. It is enough to take $U = L^t(\Gamma)$, $Z = C(X)$, J the functional to minimize, G the mapping that associates to each control the corresponding state, which is differentiable (Theorem 3.3), K the convex subset of $L^t(\Gamma)$, and C the convex subset of $Z = C(X)$ with nonempty interior. Theorem 5.2 states the existence of $\bar{\lambda}$ and $\bar{\mu}$ satisfying (5.9) and (5.12). Now let us take $\bar{y} = y_{\bar{u}}$ and $\bar{p} \in W^{1,s}(\Omega)$, the unique solution of (5.11); see Theorem 4.3. Then it remains to prove inequality (5.13), which is accomplished by using the corresponding inequality (5.3). For this, it is enough to establish the identity

$$(5.15) \quad \bar{\lambda} J'(\bar{u}) \cdot v + \langle [DG(\bar{u})]^* \bar{\mu}, v \rangle = \int_{\Gamma} (\bar{p} + \bar{\lambda} N |\bar{u}|^{\sigma-2} \bar{u}) v dm_{\Gamma}(x) \quad \forall v \in L^t(\Gamma).$$

Let $z \in H^1(\bar{\Omega}) \cap C(\bar{\Omega})$ be the solution of the Neumann problem (3.4) with $u = \bar{u}$, and let us take $\{v_k\} \subset D(\Gamma)$ converging to v in $L^t(\Gamma)$ and $z_k \in W^{2,r}(\Omega)$ satisfying

$$\begin{aligned} Az_k + \phi'(\bar{y})z_k &= 0 & \text{in } \Omega, \\ \partial_{\nu_A} z_k &= v_k & \text{on } \Gamma. \end{aligned}$$

Then it follows from Theorem 3.1 that $\{z_k\}$ converges to z uniformly in $\bar{\Omega}$ and strongly in $H^1(\Omega)$ toward z . Hence it follows from relation (4.6) applied to the operator $A^* + \phi'(\bar{y})$ (where $q = \gamma(\bar{p})$, as was proved there), taking $\mu_{\Omega} = \bar{\lambda}(\bar{y} - y_d) + \bar{\mu}_{\Omega}$ and $\mu_{\Gamma} = \bar{\mu}_{\Gamma}$, that

$$\begin{aligned} \bar{\lambda} \int_{\Omega} (\bar{y} - y_d) z dx + \int_{\bar{\Omega}} z d\bar{\mu} &= \lim_{k \rightarrow \infty} \left(\bar{\lambda} \int_{\Omega} (\bar{y} - y_d) z_k dx + \int_{\bar{\Omega}} z_k d\bar{\mu} \right) \\ &= \lim_{k \rightarrow \infty} \left(\bar{\lambda} \int_{\Omega} (\bar{y} - y_d) z_k dx + \int_{\Omega} z_k d\bar{\mu}_{\Omega} + \int_{\Gamma} z_k d\bar{\mu}_{\Gamma} \right) \\ &= \lim_{k \rightarrow \infty} \left(\int_{\Omega} \bar{p} (Az_k + \phi'(\bar{y})z_k) dx + \int_{\Gamma} \bar{p} \partial_{\nu_A} z_k dm_{\Gamma}(x) \right) \\ &= \lim_{k \rightarrow \infty} \int_{\Gamma} \bar{p} v_k dm_{\Gamma}(x) = \int_{\Gamma} \bar{p} v dm_{\Gamma}(x). \end{aligned}$$

Therefore (5.15) is deduced from the above relation and the equality

$$\begin{aligned} &\bar{\lambda} J'(\bar{u}) \cdot v + \langle [DG(\bar{u})]^* \bar{\mu}, v \rangle \\ &= \bar{\lambda} \int_{\Omega} (\bar{y} - y_d) z dx + \bar{\lambda} N \int_{\Gamma} |\bar{u}|^{\sigma-2} \bar{u} v dm_{\Gamma}(x) + \int_{\bar{\Omega}} z d\bar{\mu}. \end{aligned}$$

The qualification ($\bar{\lambda} = 1$) under the Slater condition (5.14) follows directly from Theorem 5.2. \square

Remark 3. If ϕ is a linear function, then the Slater condition (5.14) becomes

$$\exists u_0 \in K \quad \text{such that } y_{u_0} \in \overset{\circ}{C}.$$

Thus, assuming ϕ linear,

$$(5.16) \quad C = \{z \in C(X) : |z(x)| \leq \delta \quad \forall x \in X\},$$

and $\delta_0 > 0$ such that (P) has a feasible control for this value, then the Slater condition is satisfied for every $\delta > \delta_0$.

This condition allows us to deduce the optimality system (5.10)–(5.13) with $\bar{\lambda} = 1$. It is possible, however, to sometimes deduce that $\bar{\lambda} = 1$ without proving the Slater condition. For example, if C and δ_0 are defined as above, with ϕ not necessarily linear, then the optimality system is verified with $\bar{\lambda} = 1$ for almost all $\delta \in [\delta_0, +\infty)$; see Remark 2.

Remark 4. Let us suppose that C is given as above, and let be $\bar{\mu} = \bar{\mu}_\Omega + \bar{\mu}_\Gamma$; then $\bar{\mu}$ has a Jordan decomposition $\bar{\mu} = \bar{\mu}^+ - \bar{\mu}^-$ in such a way that $\bar{\mu}^+$ is concentrated in the Borel set X^+ and $\bar{\mu}^-$ is concentrated in X^- , where

$$X^+ = \{x \in X : \bar{y}(x) = +\delta\} \quad \text{and} \quad X^- = \{x \in X : \bar{y}(x) = -\delta\}.$$

In particular, if the equality $|\bar{y}(x)| = \delta$ is satisfied at a finite set of points $\{x_j\}_{j=1}^m$, then we have that

$$\bar{\mu} = \sum_{j=1}^m \lambda_j \delta_{x_j},$$

where $\lambda_j \in \mathbb{R}$ and δ_{x_j} is the Dirac measure concentrated at x_j . Furthermore, $\lambda_j \geq 0$ if $\bar{y}(x_j) = +\delta$ and $\lambda_j \leq 0$ if $\bar{y}(x_j) = -\delta$; see Casas [7].

Remark 5. From (5.13) we can deduce some qualitative properties of the optimal control \bar{u} . For example, if $N = 0$ or $\bar{\lambda} = 0$ and

$$K = \{u \in L^\infty(\Gamma) : a \leq u(x) \leq b \quad \text{a.e. } [m_\Gamma] \ x \in \Gamma\},$$

then \bar{u} has a behavior of bang-bang type. More precisely, $\bar{u}(x) = a$ if $\bar{p}(x) > 0$ and $\bar{u}(x) = b$ if $\bar{p}(x) < 0$. However, if $N\bar{\lambda} \neq 0$, then \bar{u} can possess additional regularity properties. Thus if there is not any constraint on the control, then from (5.13) it follows that

$$\bar{u}(x) = \frac{-1}{(N\bar{\lambda})^{1/(\sigma-1)}} |\bar{p}(x)|^{(2-\sigma)/(\sigma-1)} \bar{p}(x),$$

where σ is greater than or equal to $t > n - 1$.

Hence \bar{u} is continuous in the points where \bar{p} is continuous. If C is given by (5.16) and we suppose that $y_d \in L^r(\Omega)$, then \bar{p} is continuous in $\Gamma \setminus \text{support}(\bar{\mu})$, which follows from (5.11). Therefore \bar{u} is continuous in the points where the state constraint is not active.

If K is the set of positive controls, the situation is very similar, because, in this case,

$$\bar{u}(x) = \max \left\{ 0, \frac{-1}{(N\bar{\lambda})^{1/(\sigma-1)}} |\bar{p}(x)|^{(2-\sigma)/(\sigma-1)} \bar{p}(x) \right\}.$$

When $K = \{u \in L^\infty(\Gamma) : a \leq u(x) \leq b \text{ a.e. } [m_\Gamma] x \in \Gamma\}$, then σ can be taken equal to 2 and we have that

$$\bar{u}(x) = \text{Proj}_{[a,b]} \left(\frac{-1}{N\lambda} \bar{p}(x) \right).$$

Thus we can deduce again continuity of \bar{u} in the points where the state constraint is not active. Moreover, since the mapping $g : R \rightarrow R$ defined by

$$g(t) = \text{Proj}_{[a,b]} \left(\frac{-t}{N\lambda} \right)$$

is Lipschitz and $\gamma(\bar{p}) \in W^{1/r,s}(\Gamma)$, then $\bar{u} \in W^{1/r,s}(\Gamma)$, also.

REFERENCES

- [1] F. ABERGEL AND R. TEMAM, *Optimality conditions for some non qualified problems of distributed control*, SIAM J. Control Optim., 27 (1989), pp. 1–12.
- [2] A. BERMÚDEZ AND A. MARTÍNEZ, *An optimal control problem with state constraints related to the sterilization of canned foods*, to appear.
- [3] J. BONNANS AND E. CASAS, *Contrôle de systèmes non linéaires comportant des contraintes distribuées sur l'état*, Tech. Rep. 300, INRIA Rocquencourt, May 1984.
- [4] ———, *Contrôle de systèmes elliptiques semilinéaires comportant des contraintes sur l'état*, in Nonlinear Partial Differential Equations and Their Applications, Vol. 8, Collège de France Seminar, H. Brezis and J. Lions, eds., Longman Scientific & Technical, New York, 1988, pp. 69–86.
- [5] ———, *Optimal control of semilinear multistate systems with state constraints*, SIAM J. Control Optim., 27 (1989), pp. 446–455.
- [6] E. CASAS, *Quelques problèmes de contrôle avec contraintes sur l'état*, C.R. Acad. Sci. Paris, t. 296 (1983), pp. 509–512.
- [7] ———, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.
- [8] ———, *Optimality conditions and numerical approximations for some optimal design problems*, Control Cybernet., 19 (1990), pp. 73–91.
- [9] ———, *Optimal control in coefficients with state constraints*, Appl. Math. Optim., 26 (1992), pp. 21–37.
- [10] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, London, Melbourne, 1985.
- [11] J. LIONS, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.
- [12] E. LUNEVILLE, *Simulation et contrôle de la trempe superficielle par laser*, Tech. Rep. 236, Ecole Nationale Supérieure de Techniques Avancées, October 1989.
- [13] U. MACKENROTH, *Convex parabolic control problems with pointwise state constraints*, J. Math. Anal. Appl., 87 (1982), pp. 256–277.
- [14] ———, *On some elliptic optimal control problems with state constraints*, Optimization, 17 (1986), pp. 595–607.
- [15] J. NEČAS, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Editeurs Academia, Prague, 1967.
- [16] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, London, 1970.
- [17] G. STAMPACCHIA, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

SECOND-ORDER SUFFICIENT OPTIMALITY CONDITIONS FOR A CLASS OF NONLINEAR PARABOLIC BOUNDARY CONTROL PROBLEMS*

H. GOLDBERG[†] AND F. TRÖLTZSCH[†]

Abstract. In this paper sufficient second-order optimality conditions are established for parabolic boundary control problems with nonlinear boundary condition and constraints on the control and the state. The main idea is to extend the known theory for systems governed by ordinary differential equations to the case of partial differential equations. This is performed by means of a semigroup approach and a two-norm technique. The verification of the second-order conditions is discussed.

Key words. optimal control, nonlinear parabolic equation, second-order condition, sufficient optimality condition, semigroup

AMS subject classifications. 49K20, 49K27, 90C48, 90C31

1. Introduction. This paper is a further contribution to the theory of optimality conditions for optimal control problems with distributed parameters. The control system under consideration is governed by a semilinear parabolic equation; hence the control problem belongs to the class of nonconvex optimization problems. In contrast to parabolic control problems with convex objective functional and linear equations, where the list of references on optimality conditions is very extensive, only a few investigations have been devoted to the case of nonlinear parabolic equations. We only mention Friedman [9], Sachs [22], Schmidt [23], and Tröltzsch [24], whose papers are closely related to our work. They are concerned mainly with first-order necessary optimality conditions in the form of “local” maximum principles. Another group of publications is devoted to generalizations of the Pontrjagin maximum principle, which avoids the linearization with respect to the control (being typical for “local” maximum principles). We refer to Fattorini [8], [6], and von Wolfersdorf [29].

First-order optimality conditions are very useful to derive structural properties of optimal controls such as bang-bang theorems and their generalizations (see, for instance, [24]). However, they are lacking in the sufficiency for nonconvex problems. Therefore, their application to the numerical analysis of optimal control problems is limited mainly to the convex case, where the strong convergence of sequences of optimal control of (FEM-) approximations of the control problems can be shown. A number of papers is concerned with such investigations, for instance by Lasiecka [15], [17], Knowles [13], Alt and Mackenroth [1], Malanowski [19], and others.

In nonconvex problems, sufficient second-order conditions at the optimal point are a substitute for convexity. The theory of sufficient second-order conditions for twice differentiable extremal problems in function spaces is known to be more rich and interesting than that for problems in finite-dimensional spaces. This is due to the so-called two-norm discrepancy, expressing the noncompatibility of the norms needed for second-order optimality conditions. This difficulty was resolved successfully by Ioffe [12] and Maurer [20]. Based on these general results a satisfactory theory of sufficient second-order conditions and its application to nonlinear optimal control problems governed by ordinary differential equations was worked out. Our paper seeks to

* Received by the editors April 17, 1991; accepted for publication (in revised form) March 20, 1992.

[†] Fachbereich Mathematik, Technische Universität Chemnitz, PSF 964, O-9010 Chemnitz, Germany.

contribute to an analogous theory of second-order sufficient optimality conditions for control problems governed by semilinear parabolic initial-boundary value problems with constraints on the control and the state. We continue our investigations in [10], where a control problem for the one-dimensional heat equation without state constraints was considered. For a higher-dimensional version we refer to [11]. A first application of these results to the numerical approximations of the corresponding problem is contained in Tröltzsch [27].

The extension to higher-dimensional problems is based on a semigroup approach. We rely heavily upon recent results by Amann [3], [2], Fattorini [7], Lasiecka [16], and others. It should be emphasized that, in contrast to the treatment of control problems for ordinary differential equations, L_2 controls are not transformed to continuous state functions (even if the control appears only linearly). In view of this, a two-norm technique is indispensable for a satisfactory handling of the problems (at least, if continuity of the state is needed to define the objective functional or the state constraints).

In this paper we use the following notation.

Let X, Y be real Banach spaces. Then $\mathcal{L}(X, Y)$ is the space of linear continuous operators from X to Y , $\mathcal{L}(X) = \mathcal{L}(X, X)$. X^* denotes the dual space to X , $A^* \in \mathcal{L}(Y^*, X^*)$ the adjoint operator to $A \in \mathcal{L}(X, Y)$. By $(\cdot, \cdot)(D)$ the pairing between $L_p(D)$ and $L_q(D)$, $1/p + 1/q = 1$, $1 \leq p < \infty$, is denoted (if p, q are not specified, then this sign stands simply for integration on D). For $\Omega \in \mathbb{R}^n$ we shall write $\Omega_T = [0, T] \times \Omega$. Moreover, we will work in the following spaces:

$$\begin{aligned} X_p &= U_p &= L_p(0, T; L_p(\Gamma)), & 1 \leq p < \infty \\ X_\infty &= C([0, T], C(\Gamma)) \\ U_\infty &= L_\infty((0, T) \times \Gamma) \\ W_p^\sigma(\Omega) &- \text{Sobolev-Slobodeckij-space} \end{aligned}$$

2. Formulation of the control problem. We consider the optimal control problem to minimize

$$\int_{\Omega} \varphi(x, w(T, x)) dx + \int_0^T \int_{\Omega} \psi(t, x, w(t, x)) dx dt + \int_0^T \int_{\Gamma} \chi(t, x, w(t, x), u(t, x)) dS_x dt$$

subject to the equation of state

$$(2.1) \quad \begin{aligned} w_t(t, x) &= (\Delta_x - 1)w(t, x) && \text{on } (0, T] \times \Omega \\ w(0, x) &= w_0(x) && \text{on } \Omega \\ \frac{\partial w}{\partial n}(t, x) &= b(t, x, w(t, x), u(t, x)) && \text{on } (0, T] \times \Gamma, \end{aligned}$$

where u is looked upon as a *control* subject to

$$(2.2) \quad u_1(t, x) \leq u(t, x) \leq u_2(t, x) \quad \text{a.e. on } (0, T] \times \Gamma.$$

Furthermore, we are able to include state constraints:

$$(2.3) \quad \int_{\Omega} \Phi_i(x) w(t, x) dx \leq c_i(t) \quad \text{on } [0, T], \quad i = 1, \dots, k.$$

The state $w \in C([0, T], W_p^\sigma(\Omega))$ of the control system is defined below as mild solution for (2.1) and the control u is taken from $L_\infty((0, T) \times \Gamma)$. In the problem the following

quantities occur: $\Omega \in \mathbb{R}^n$, $n \geq 2$, is a bounded domain with C^∞ -boundary Γ , $T > 0$ is a fixed time. $\Phi_i \in W_p^\sigma(\Omega)$, $i = 1, \dots, k$, $w_0 \in W_p^\sigma(\Omega)$, $u_1, u_2 \in L_\infty((0, T) \times \Gamma)$ with $u_1(t, x) < u_2(t, x)$ on $[0, T] \times \Gamma$, and $c_i \in C[0, T]$, $i = 1, \dots, k$, are real-valued functions. Moreover, $\varphi : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$, $\psi : [0, T] \times \Omega \times \mathbb{R} \rightarrow \mathbb{R}$, and $\chi, b : [0, T] \times \Gamma \times \mathbb{R}^2 \rightarrow \mathbb{R}$ are nonlinear functions. They are supposed for convenience to be twice continuously differentiable on their domains (although this could be weakened partially to natural measurability assumptions with respect to (t, x)). By $\partial/\partial n$ we denote the outward normal derivative at Γ , dS_x is the surface measure on Γ .

Remark 1. The choice of the differential operator $\Delta_x - I$ is only for technical reasons to make the corresponding elliptic Neumann problem uniquely solvable. By the simple transformation $\tilde{w}(t, x) = e^{-t}w(t, x)$ the case Δ_x can be transformed back to our problem (with redefined nonlinear functions). Moreover, the theory works analogously for more general uniformly elliptic differential operators with C^∞ -coefficients.

The function $b = b(t, x, w, u)$ defines a Nemytskij operator B by

$$B(w, u)(t, x) = b(t, x, w(t, x), u(t, x))$$

from $C([0, T] \times \Gamma) \times L_\infty((0, T) \times \Gamma)$ to $L_\infty((0, T) \times \Gamma)$. B is twice continuously Fréchet differentiable owing to the assumptions on b . However, we shall define B in slightly changed spaces.

It is obvious that $C([0, T] \times \Gamma) = C([0, T], C(\Gamma)) = X_\infty$. Moreover $L_p(0, T; L_p(\Gamma)) = L_p((0, T) \times \Gamma)$, $1 \leq p < \infty$ (each equivalence class of functions of a space can be represented by one, belonging to the other space), but only

$$L_\infty(0, T; L_\infty(\Gamma)) \subset L_\infty((0, T) \times \Gamma)$$

(cf. the simple example given by Fattorini [6]). Therefore, in what follows we regard B as an operator from $X_\infty \times U_\infty$ to X_p . Clearly B remains twice Fréchet differentiable in this more general setting. We indicate the trace operator by τ .

DEFINITION 1 (cf. [2] and [3]). Any $w \in W_p^\sigma(\Omega)$ satisfying

$$(2.4) \quad w(t) = S(t)w_0 + \int_0^t AS(t-s)NB(\tau w, u)(s) ds, \quad t \in [0, T],$$

is called a *mild solution* of (2.1). Here $A : L_p(\Omega) \supset D(A) \rightarrow L_p(\Omega)$ is defined by

$$D(A) = \left\{ w \in W_p^2(\Omega) : \frac{\partial w}{\partial n} = 0 \right\}, \quad Aw = -\Delta w + w,$$

S is the semigroup generated by $-A$ in $L_p(\Omega)$, and the Neumann operator $N : L_p(\Gamma) \rightarrow W_p^\sigma(\Omega)$ assigns to g the solution w of $\Delta w - w = 0$, $\partial w/\partial n = g$. The parameters p and σ are fixed subject to $p > n + 1$ and

$$(2.5) \quad \frac{n}{p} < \sigma < 1 + \frac{1}{p}.$$

We should note that (2.5) implies $W_p^\sigma(\Omega) \hookrightarrow C(\bar{\Omega})$ and $W_p^{\sigma-(1/p)}(\Gamma) \hookrightarrow C(\Gamma)$; hence $\tau w \in X_\infty$.

Completely analogous, operators $A_r, S_r(t)$, and N_r are introduced substituting $r \in (1, \infty)$ for p in Definition 1. Thus we have $A = A_p, S = S_p, N = N_p$.

The properties of the solution of (2.4) have been discussed extensively by Amann; we refer, for instance, to [2] and [3]. It was shown that a mild solution w is also a weak solution (cf. [2]). For the case of control problems see also Tröltzsch [28]: There is a sufficiently small $T > 0$ such that for all $u \in U_{ad}$ satisfying (2.2) a unique solution $w \in C([0, T], W_p^\sigma(\Omega))$ of (2.4) exists. The key to this result is that $A_r S_r(t) N_r$ is a continuous operator from $L_p(\Gamma)$ to $W_p^\sigma(\Omega)$ for $t > 0$ together with the estimate

$$(2.6) \quad \|A_r S_r(t) N_r\|_{L_r(\Gamma) \rightarrow W_p^\sigma(\Omega)} \leq ct^{-(1-(\sigma'-\sigma)y/2)},$$

for all $0 < \sigma < \sigma' < 1 + 1/r$ derived by Amann [3]. We assume throughout this paper that $T > 0$ meets this requirement. Often we can proceed on the assumption $T = \infty$, we mention only [23], which considered several practical important types of nonlinear boundary conditions.

The presence of the state-constraint (2.3) essentially complicates the treatment of our nonlinear optimal control problem. This difficulty can be resolved by embedding the problem into a general class of nonlinear programs in Banach spaces with equality and inequality constraints. Therefore, it is natural and necessary to invoke the corresponding extensive theory of optimality conditions. In view of this, we now convert the control problem into a mathematical programming problem.

We want to minimize

$$f(w, u) := f^1(w(T)) + f^2(w) + f^3(w, u),$$

where

$$\begin{aligned} f^1 &: C(\overline{\Omega}) \rightarrow \mathbb{R}, \\ f^1(w) &= \int_{\Omega} \varphi(x, w(x)) \, dx \\ f^2 &: C([0, T], C(\overline{\Omega})) \rightarrow \mathbb{R}, \\ f^2(w) &= \int_0^T \int_{\Omega} \psi(t, x, w(t, x)) \, dx \, dt \\ f^3 &: X_\infty \times L_\infty((0, T) \times \Gamma) \rightarrow \mathbb{R}, \\ f^3(w, u) &= \int_0^T \int_{\Gamma} \chi(t, x, w(t, x), u(t, x)) \, dS_x \, dt. \end{aligned}$$

The state constraints can be formalized by linear operators G_i ,

$$(G_i w)(t) = \int_{\Omega} \Phi_i(x) w(t, x) \, dx,$$

being continuous from $C([0, T], C(\overline{\Omega}))$ to $C[0, T]$. After introducing the operators

$$\begin{aligned} (Lz)(t) &= \int_0^t AS(t-s)Nz(s) \, ds, \\ (Kz)(t) &= (\tau Lz)(t), \\ \Lambda z &= (Lz)(T), \end{aligned}$$

the new state function $v(t) = \tau w(t)$, and $d(t) = S(t)w_0$, we can formulate our control problem as

$$(P) \quad \begin{aligned} & f^1(d(T) + \Lambda B(v, u)) + f^2(d + LB(v, u)) + f^3(v, u) = \min! \\ & v = \tau d + KB(v, u), \\ & G_i(d + LB(v, u)) \leq c_i, \quad i = 1, \dots, k, \\ & u \in C. \end{aligned}$$

3. First-order necessary optimality conditions. A pair $(v, u) \in X_\infty \times U_\infty$ satisfying all constraints of (P) is said to be *admissible*. In what follows let the admissible (v^o, u^o) be locally optimal for (P). Then u^o is said to be an *optimal control*. That means $F(v^o, u^o) \leq F(v, u)$ for all (v, u) being admissible and contained in a sufficiently small neighbourhood of (v^o, u^o) in the space $X_\infty \times U_\infty$. If u is sufficiently close to u^o , then so is v to v^o . Hence local optimality can also be formulated in terms of u only. For computing the Fréchet derivatives of the nonlinear functionals and operators under consideration we need the first- and second-order derivatives of φ , χ , ψ , b at the optimal pair. We indicate them by corresponding subscripts and omit the dependence on v^o , u^o , w^o . For instance,

$$\begin{aligned} \psi_w(t, x) &= \frac{\partial \psi}{\partial w}(t, x, w^o(t, x)), \\ \psi_{ww}(t, x) &= \frac{\partial^2 \psi}{\partial w^2}(t, x, w^o(t, x)). \end{aligned}$$

In this way, the first-order Fréchet derivatives admit the following form:

$$\begin{aligned} (f^1)'(w^o(T))w &= \int_{\Omega} \varphi_w(x)w(x) dx, \quad (f^2)'(w^o)w = \int_0^T \int_{\Omega} \psi_w(t, x)w(t, x) dx dt \\ (f^3)'(v^o)h &= \int_0^T \int_{\Gamma} \chi_w(t, x)v(t, x) dS_x dt + \int_0^T \int_{\Omega} \chi_u(t, x)z(t, x) dS_x dt, \\ (h = (v, z) \in X_\infty \times U_\infty), \end{aligned}$$

and

$$B'(v^o, u^o)h = B_v v + B_u z,$$

where

$$(B_v v)(t, x) = b_w(t, x)v(t, x), \quad (B_u z)(t, x) = b_u(t, x)z(t, x).$$

The functions φ_w , ψ_w , χ_w , χ_u , b_w , and b_u are bounded and measurable on their domains. Hence the linear functionals $(f^i)'$ and the linear operators B_v , B_u extend continuously to all corresponding L_p -spaces (p according to Definition 1). In the following we regard these extensions and use the same notation as before. In doing so, we have $(f^1)' \in L_p(\Omega)^*$, $(f^2)' \in X_p^*$, $(f^3)' \in X_p^* \times X_p^*$, $B_u, B_v \in \mathcal{L}(X_p)$. It should be emphasized that we first determine the derivative in $X_\infty \times U_\infty$. Only the derivatives, after having been computed, are extended to $X_p \times U_p$. The second-order Fréchet derivative of B at (v^o, u^o) is given by

$$(B''(v^o, u^o)[h, h])(t, x) = h(t, x)^T b''(t, x)h(t, x),$$

where $h(t, x)^T = (v(t, x), u(t, x))$ and

$$b''(t, x) = \begin{pmatrix} b_{ww}(t, x) & b_{wu}(t, x) \\ b_{uw}(t, x) & b_{uu}(t, x) \end{pmatrix}$$

(partial derivatives taken at $(t, x, v^o(t, x), u^o(t, x))$). To formulate the optimality conditions we introduce the *Lagrange function*

$$\begin{aligned} \mathcal{L}(v, u; y, \lambda) &= F(v, u) + \int_0^T \int_{\Gamma} (v - \tau d - KB(v, u))(t, x) y(t, x) dS_x dt \\ &+ \sum_{i=1}^k \int_0^T G_i(d + LB(v, u))(t) d\lambda_i(t). \end{aligned}$$

The operators K , L , and Λ have their range in spaces of continuous abstract functions. Embedding the range spaces into corresponding L_p -spaces we can regard them as operators from L_p to L_p . We do so in what follows; i.e., we define K , L , and Λ as operators between the following spaces: $K : X_p \rightarrow X_p$, $L : X_p \rightarrow L_p(0, T; L_p(\Omega))$, $\Lambda : X_p \rightarrow L_p(\Omega)$. Therefore, $K^* : X_q \rightarrow X_q$, $L^* : L_q(0, T; L_q(\Omega)) \rightarrow X_q$, $\Lambda^* : X_q \rightarrow L_q(\Omega)$ ($1/p + 1/q = 1$). The kernels of these operators are regarded in L_p -spaces, too: $AS(t)N : L_p(\Gamma) \rightarrow L_p(\Omega)$, $\tau AS(t)N : L_p(\Gamma) \rightarrow L_p(\Gamma)$, $t > 0$. Their adjoint operators can be determined by a simple integration by parts, cf. [24]: $(AS(t)N)^* = \tau S_q(t) : L_q(\Omega) \rightarrow L_q(\Gamma)$, $(\tau AS(t)N)^* = \tau A_q S_q(t) N_q : L_q(\Gamma) \rightarrow L_q(\Gamma)$. Thus

$$\begin{aligned} (K^*y)(t) &= \int_t^T \tau A_q S_q(s-t) N_q y(s) ds, \\ (L^*w)(t) &= \int_t^T \tau S_q(s-t) w(s) ds, \\ (\Lambda^*\varphi)(t) &= \tau S_q(T-t) \varphi. \end{aligned}$$

For the proof of a Lagrange multiplier rule we need a certain regularity condition.

DEFINITION 2. The pair $(v^o, u^o) \in X_\infty \times U_\infty$ is said to be *regular* for (P), if there exists a pair $(\bar{v}, \bar{u}) \in X_\infty \times C$ such that

$$(3.1) \quad \bar{v} - v^o = K(B_v(\bar{v} - v^o) + B_u(\bar{u} - u^o))$$

$$(3.2) \quad (G_i(w^o) + G_i L(B_v(\bar{v} - v^o) + B_u(\bar{u} - u^o)))(t) < c_i(t)$$

on $[0, T]$, $i = 1, \dots, k$, where $w^o := d + LB(v^o, u^o)$.

THEOREM 3.1. Suppose that (v^o, u^o) is a regular locally optimal solution of the optimal control problem (P). Then there are $y \in L_q(0, T; L_q(\Gamma))$ and monotone non-decreasing $\lambda_i \in NBV[0, T]^1$ such that

$$(3.3) \quad \mathcal{L}_v(v^o, u^o; y, \lambda) = 0$$

$$(3.4) \quad \mathcal{L}_u(v^o, u^o; y, \lambda)(u - u^o) \geq 0 \quad \forall u \in C$$

$$(3.5) \quad \int_0^T (G_i(w^o) - c_i)(t) d\lambda_i(t) = 0, \quad i = 1, \dots, k,$$

¹ Space of functions of bounded variation with the normalization condition $\lambda_i(T) = 0$.

where $\mathcal{L}_v, \mathcal{L}_u$ denote the partial Fréchet derivatives of L at (v^o, u^o) in the space $X_\infty \times U_\infty$, $\lambda = (\lambda_1, \dots, \lambda_k)$.

We sketch the proof only briefly. The underlying two-space technique is derived in a more elegant way in [24], [25]:

Proof. As a (nontrivial) conclusion of the regularity condition we know that (v^o, u^o) is the solution of the *linearized control problem*

$$\begin{aligned} F_v(v^o, u^o)v + F_u(v^o, u^o)u &= \min!, \\ v &= K(B_v v + B_u(u - u^o)), \\ G_i(w^o) + G_i L(B_v v + B_u(u - u^o)) &\leq c_i, \quad i = 1, \dots, k, \\ u &\in C. \end{aligned}$$

In the next step we extend the space $X_\infty \times U_\infty$ to $X_p \times U_p$; i.e., we look for all solutions of this problem in $X_p \times U_p$. As K and L map X_p into spaces of continuous functions, the linearized admissible set remains unchanged. Moreover, continuity and extension properties of $B_v, B_u, (f^i)'$, $i = 1, 2, 3$, imply that F_v and F_u can be continuously extended to X_p and U_p . In view of this, we can assume $F_v \in X_p^*, F_u \in U_p^*$. On the other hand, the linearized problem in $X_p \times U_p$ satisfies the regularity condition at (\bar{v}, \bar{u}) , too. Therefore, a Lagrange multiplier rule is valid: There exist $y \in X_q, \lambda_i \in NBV[0, T]$ such that

$$(3.6) \quad \begin{aligned} &(\Lambda B_v v, \varphi_w)(\Omega) + (LB_v v, \psi_w)(\Omega_T) + (\chi_w, v)(\Gamma_T) \\ &+ (v - KB_v v, y)(\Gamma_T) + \sum_{i=1}^k \int_0^T (G_i LB_v v)(t) d\lambda_i(t) = 0 \end{aligned}$$

for all $v \in X_p$,

$$(3.7) \quad \begin{aligned} &(\Lambda B_u(u - u^o), \varphi_w)(\Omega) + (LB_u(u - u^o), \psi_w)(\Omega_T) \\ &+ ((\chi_u - KB_u)(u - u^o), y)(\Gamma_T) + \sum_{i=1}^k \int_0^T (G_i LB_u(u - u^o))(t) d\lambda_i(t) \geq 0 \end{aligned}$$

for all $u \in C$, and the complementary slackness condition (3.5) holds. Writing down \mathcal{L}_v and \mathcal{L}_u we see that (3.6) and (3.7) are equivalent to (3.3) and (3.4). \square

The concrete expression of $(G_i L)^*$ is derived in the following lemma.

LEMMA 3.2. For $(G_i L)^* : NBV[0, T] \rightarrow L_q(0, T; L_q(\Gamma))$

$$(L^* G_i^* \Phi_i)(t) = \int_t^T \tau S_p(s - t) \Phi_i d\lambda_i(s)$$

holds. The function $S_p(s - t)\Phi_i$ belongs to $C(D, C(\bar{\Omega}))$, where $D = \{(t, s) | 0 \leq t \leq s \leq T\}$.

Proof. We have

$$(AS(s - t)N)^* \Phi_i = \tau S_q(s - t) \Phi_i = \tau S_p(s - t) \Phi_i,$$

as $\Phi_i \in W_p^\sigma(\Omega)$. This follows by means of Pazy [21, Chap. 4, Thm. 5.5]. Moreover, it is known that $S_p(t)$ restricts to a strongly continuous semigroup on $W_p^\sigma(\Omega)$, cf. [3]; hence

$S_p(s-t)\Phi_i$ is a continuous abstract function on D with values in $W_p^\sigma(\Omega) \hookrightarrow C(\overline{\Omega})$. This yields the second assertion of the lemma. Thus

$$\begin{aligned} \int_0^T (G_i Lz)(t) d\lambda_i(t) &= \int_0^T (\Phi_i, \int_0^t AS(t-s)Nz(s) ds)(\Omega) d\lambda_i(t) \\ &= \int_0^T \int_0^t (\tau S_p(t-s)\Phi_i, z(s))(\Gamma) ds d\lambda_i(t) \\ &= \int_0^T \int_t^T (\tau S_p(s-t)\Phi_i d\lambda_i(s), z(t))(\Gamma) dt \\ &= \int_0^T ((G_i L\Phi_i)^*(t), z(t))(\Gamma) dt, \end{aligned}$$

where the abstract Riemann–Stieltjes integral exists due to the continuity of $\tau S_p(s-t)\Phi_i$. \square

Using in (3.6) and (3.7) the (L_p-) adjoint of the linear operators we arrive at

$$\begin{aligned} \mathcal{L}_v(v^o, u^o; y, \lambda)v &= \left(v, y + B_v^* \left\{ -K^*y + \Lambda^*\varphi_w + \sum_{i=1}^k L^*G_i^*\lambda_i \right\} + \chi_w \right) (\Gamma_T) \\ (3.8) \quad &=: \int_0^T \int_{\Gamma} v(t, x) \mathcal{L}_v(t, x) dS_x dt, \end{aligned}$$

$$\begin{aligned} \mathcal{L}_u(v^o, u^o; y, \lambda)u &= \left(u, B_u^* \left\{ -K^*y + \Lambda^*\varphi_w + \sum_{i=1}^k L^*G_i^*\lambda_i \right\} + \chi_u \right) (\Gamma_T) \\ (3.9) \quad &=: \int_0^T \int_{\Gamma} u(t, x) \mathcal{L}_u(t, x) dS_x dt. \end{aligned}$$

From the optimality conditions,

$$y(t) = -b_w(t, \cdot) \left\{ -K^*y + \Lambda^*\varphi_w + L^*\psi_w + \sum_{i=1}^k L^*G_i^*\lambda_i \right\} (t) - \chi_w(t, \cdot);$$

hence, after inserting the expressions for the adjoint operators,

$$\begin{aligned} y(t, \cdot) &= -b_w(t, \cdot) \left\{ - \int_t^T \tau A_q S_q(s-t) N_q y(s) ds + \tau S_q(T-t) \varphi_w \right. \\ (3.10) \quad &\quad \left. + \int_t^T \tau S_q(s-t) \psi_w(s) ds + \sum_{i=1}^k \int_t^T \tau S_p(s-t) \Phi_i d\lambda_i(s) \right\} - \chi_w(t, \cdot). \end{aligned}$$

This may be defined as an adjoint equation. However, it is more convenient to intro-

duce

$$(3.11) \quad p(t, \cdot) = \left\{ - \int_t^T \tau A_q S_q(s-t) N_q y(s) ds + \tau S_q(T-t) \varphi_w \right. \\ \left. + \int_t^T \tau S_q(s-t) \psi_w(s) ds + \sum_{i=1}^k \int_t^T \tau S_p(s-t) \Phi_i d\lambda_i(s) \right\}$$

as a new adjoint state. This function can be interpreted as the mild solution of an adjoint parabolic initial boundary value problem (cf. [26]).

4. Second-order sufficient optimality conditions. In what follows let (v^o, u^o) be an admissible pair for (P). The set $M(v^o, u^o)$ consisting of all elements $(k, z) \in X_\infty \times U_\infty$ with

$$k = K(B_v k + B_u z), \quad G_i w^o + G_i L(B_v k + B_u z) \leq c_i, \quad i = 1, \dots, k,$$

$z = \lambda(u - u^o)$, $\lambda \geq 0$, $u \in C$, is said to be the *linearized set at (v^o, u^o)* . By $r_2^{\mathcal{L}}(h)$ we denote the second-order remainder term of \mathcal{L} at (v^o, u^o) in the direction $h = (v - v^o, u - u^o) \in X_\infty \times U_\infty$:

$$(4.1) \quad r_2^{\mathcal{L}}(h) = \mathcal{L}(v, u) - \mathcal{L}(v^o, u^o) - \langle \mathcal{L}_v, v - v^o \rangle - \langle \mathcal{L}_u, u - u^o \rangle - \frac{1}{2} \mathcal{L}''(v^o, u^o)[h, h].$$

Moreover, we will use the following norms throughout this section: For $1 \leq \alpha \leq \infty$ we denote by $\|\cdot\|_\alpha$ the norm of U_α . The product space $X_\alpha \times U_\alpha$ will be confined with the norm

$$\|(v, u)\|_\alpha := \max(\|v\|_\alpha, \|u\|_\alpha).$$

To overcome the known “two-norm discrepancy,” which is the main difficulty to derive sufficient second-order conditions, we follow Maurer [20] and make the following assumptions.

- (A1) For all admissible (v, u) there is a pair $(k, z) = (k(v, u), z(v, u))$ belonging to the linearized set $M(v^o, u^o)$ such that for $h = (v - v^o, u - u^o)$

$$\|(k, z) - h\|_2 \|h\|_2^{-1} \rightarrow 0,$$

as $\|h\|_\infty \rightarrow 0$.

- (A2) (i) $|r_2^{\mathcal{L}}(h)| \|h\|_2^{-2} \rightarrow 0$, as $\|h\|_\infty \rightarrow 0$;
(ii) There exists $c > 0$: $|\mathcal{L}''(v^o, u^o)[(v_1, u_1), (v_2, u_2)]| \leq c \|(v_1, u_1)\|_2 \|(v_2, u_2)\|_2$
for all $(v_i, u_i) \in X_\infty \times U_\infty$, $i = 1, 2$.

Then the following assertion holds.

THEOREM 4.1 (Second order sufficient optimality condition). *Suppose that (A1), (A2) are satisfied. Let (v^o, u^o) be admissible for (P) and fulfill the first-order necessary condition (3.3)–(3.5). Suppose further the existence of a $\delta > 0$ such that*

$$(4.2) \quad \mathcal{L}''(v^o, u^o)[(k, z), (k, z)] \geq \delta \|(k, z)\|_2^2$$

for all $(k, z) \in M(v^o, u^o)$. Then there exist positive α and ϱ such that

$$(4.3) \quad F(v, u) \geq F(v^o, u^o) + \alpha \|(v - v^o, u - u^o)\|_2^2$$

for all admissible (v, u) with $\|(v - v^o, u - u^o)\|_\infty \leq \varrho$.

Proof. We sketch the essential steps of the proof, differing in some details from that given by [20].

Suppose that (v, u) is an admissible pair. Assumption (A1) implies the existence of (k, z) belonging to $M(v^o, u^o)$ and

$$v - v^o = k(v, u) + w_1(v, u), \quad u - u^o = z(v, u) + w_2(v, u),$$

where

$$\|w_1(v, u)\|_2 \|v - v^o\|_2^{-1} \rightarrow 0, \quad \|w_2(v, u)\|_2 \|u - u^o\|_2^{-1} \rightarrow 0$$

as $\|(v - v^o, u - u^o)\|_\infty \rightarrow 0$. Now we begin to estimate the objective functional:

$$F(v, u) \geq \mathcal{L}(v, u)$$

follows from the fact that the state equation is fulfilled and the state constraints are satisfied. From the Taylor expansion of \mathcal{L} and (3.3)–(3.5)

$$\begin{aligned} F(v, u) &\geq F(v^o, u^o) + \frac{1}{2} \mathcal{L}''(v^o, u^o)[(v - v^o, u - u^o), (v - v^o, u - u^o)] \\ &\quad + r_2^{\mathcal{L}}(v - v^o, u - u^o). \end{aligned}$$

Using essentially (A2)(ii) it can be shown, that (4.2) remains true, if k and z underlie a sufficiently small perturbation: There exist positive δ_0 and γ such that

$$\mathcal{L}''(v^o, u^o)[(k + w_1, z + w_2), (k + w_1, z + w_2)] \geq \delta_0 \|(k + w_1, z + w_2)\|_2^2$$

holds for

$$\|w_1(v, u)\|_2 \leq \gamma \|k\|_2 \quad \text{and} \quad \|w_2(v, u)\|_2 \leq \gamma \|z\|_2.$$

These inequalities follow from (A1), if $\|(v - v^o, u - u^o)\|_\infty \leq \varrho_1$ with a certain positive ϱ_1 . We obtain

$$F(v, u) \geq F(v^o, u^o) + \frac{1}{2} \delta_0 \|(v - v^o, u - u^o)\|_2^2 + r_2^{\mathcal{L}}(v - v^o, u - u^o).$$

(A2)(i) yields the existence of a positive ϱ_2 , such that for all admissible (v, u) with $\|(v - v^o, u - u^o)\|_\infty \leq \varrho_2$

$$|r_2^{\mathcal{L}}(v - v^o, u - u^o)| \leq \frac{\delta_0}{4} \|(v - v^o, u - u^o)\|_2^2.$$

Choosing $\varrho = \min\{\varrho_1, \varrho_2\}$ and $\alpha = \delta_0/4$ we get (4.3).

Due to (4.3), $F(v, u) > F(v^o, u^o)$ for all admissible (v, u) in a sufficiently small $X_\infty \times U_\infty$ -neighbourhood of (v^o, u^o) . \square

Remark 2. In addition to the first- and second-order condition the proof of the theorem does not invoke any other assumptions than (A1), (A2), and the differentiability properties of B and \mathcal{L} . Therefore, Theorem 2 remains true for $U_\infty := L_p(0, T; L_p(\Gamma))$ and $\|(k, z)\|_\infty := \max\{\|k\|_\infty, \|z\|_p\}$, provided that B and \mathcal{L} are twice continuously differentiable in the space $X_\infty \times U_p$. This is true for the choice

$$(4.4) \quad b(t, x, w, u) = b_1(t, x, w) + b_2(t, x, w)u$$

and

$$(4.5) \quad \chi(t, x, w, u) = \chi_1(t, x, w) + \chi_2(t, x, w)u + (w, u)\chi_3(t, x)(w, u)^T,$$

where χ_3 is a 2×2 matrix with L_∞ entries.

In the remainder of our paper we shall verify the assumptions (A1), (A2). Assumption (A2) will follow from the differentiability properties of \mathcal{L} and the special behaviour of the linear operators K , L , Λ . Assumption (A1) is implied by the regularity of (v^o, u^o) .

The second-order derivative $\mathcal{L}''(v^o, u^o)[h_1, h_2]$, $h_i = (v_i, u_i) \in X_\infty \times U_\infty$, is

$$(4.6) \quad \begin{aligned} \mathcal{L}''(v^o, u^o)[h_1, h_2] = & \int_0^T \int_\Gamma h_1(t, x)^T \{ \chi''(t, x) + p(t, x)b''(t, x) \} h_2(t, x) dS_x dt \\ & + \int_0^T \int_\Omega \psi_{ww}(t, x)(LB'h_1)(t, x)(LB'h_2)(t, x) dx dt \\ & + \int_\Omega \varphi_{ww}(x)(\Lambda B'h_1)(x)(\Lambda B'h_2)(x) dx, \end{aligned}$$

where $\chi''(t, x)$ is defined analogously to $b''(t, x)$ in §3 and $p(t, x)$ is taken from (3.11),

$$\psi_{ww}(t, x) = \psi_{ww}(t, x, w^o(t, x)), \varphi_{ww} = \varphi_{ww}(x, w^o(T, x)) \text{ and } B' = B'(v^o, u^o).$$

The corresponding computations are too lengthy to be presented here. They are along the lines of the one-dimensional case discussed in [10] and use mainly the formula for the derivative of $q(z) = Q(e + TB(z))$ (with fixed element e , linear continuous operator T):

$$\begin{aligned} q''(z^o)[h_1, h_2] = & Q''(e + TB(z^o))[TB'(z^o)h_1, TB'(z^o)h_2] \\ & + \langle q'(z^o), TB''(z^o)[h_1, h_2] \rangle. \end{aligned}$$

Now we are going to verify (A1), (A2). The key to showing (A2) is that $y(t, x)$ and thus also $p(t, x)$ is bounded and measurable on $(0, T) \times \Gamma$.

LEMMA 4.2. *The function $y(t, x)$ is bounded and measurable on $(0, T) \times \Gamma$.*

Proof. Equation (3.10) admits the form

$$(4.7) \quad y(t, \cdot) = -b_w(t, \cdot) \left\{ h(t) - \int_t^T \tau A_q S_q(s-t) N_q y(s, \cdot) ds \right\} - \chi_w(t, \cdot)$$

with

$$h(t) = \tau S_q(T-t)\varphi_w + \int_t^T \tau S_q(s-t)\psi_w(s) ds + \sum_{i=1}^k \int_t^T \tau S_q(s-t)\Phi_i d\lambda_i(s).$$

We will prove below that $h \in L_\infty((0, T) \times \Gamma)$, thus $h \in X_p$, too. It follows from Pazy [21, Chap. 4, Thm. 5.5], that the part of A_q in $L_p(\Omega)$ is A_p and the restriction of $S_q(t)$ to $L_p(\Omega)$ coincides with $S_p(t)$. Therefore, $v \in X_p$ implies

$$(4.8) \quad A_q S_q(s-t) N_q v(s) = A_p S_p(s-t) N_p v(s).$$

The real-valued function $b_w = b_w(t, x)$ is bounded and measurable on $(0, T) \times \Gamma$. Now we regard the slightly changed equation

$$\hat{y}(t, \cdot) = -b_w(t, \cdot) \left\{ h(t) - \int_t^T \tau A_p S_p(s-t) N_p \hat{y}(s, \cdot) ds \right\} - \chi_w(t, \cdot).$$

It admits a unique solution $\hat{y} \in X_p$. Moreover \hat{y} is bounded and measurable on $(0, T) \times \Gamma$, as the integral operator maps X_p in X_∞ , $h, b_w, \chi_w \in L_\infty((0, T) \times \Gamma)$. From (4.8),

$$A_q S_q(s-t) N_q \hat{y}(s) = A_p S_p(s-t) N_p \hat{y}(s);$$

hence \hat{y} solves (4.7), too. By the uniqueness of the solution to (4.7) we conclude $y = \hat{y}$; hence $y \in L_\infty((0, T) \times \Gamma)$. It remains to show that $h \in L_\infty((0, T) \times \Gamma)$. The function $z(t) = S_q(T-t)\varphi_w$ solves the parabolic problem

$$-z_t(t, x) = \Delta z(t, x) - z(t, x), \quad z(T, x) = \varphi_w(x)$$

subject to homogeneous Neumann boundary conditions, where $\varphi_w(x) = \varphi_w(x, w^\circ(T, x))$. On the other hand, $w^\circ(T, x)$ is continuous on $\bar{\Omega}$, as $w^\circ \in W_p^\sigma(\Omega) \hookrightarrow C(\bar{\Omega})$. Invoking the maximum principle, $|z(t, x)| \leq \max_{x \in \bar{\Omega}} \varphi_w(x)$. Clearly, this implies $\tau z \in L_\infty((0, T) \times \Gamma)$. The real-valued function $\psi_w(t, x)$ is bounded and measurable on $(0, T) \times \Omega$, hence $\psi_w \in L_p(0, T; L_p(\Omega))$. From [2],

$$\|S_p(s-t)\|_{L_p(\Omega) \rightarrow W_p^\sigma(\Omega)} \leq c|s-t|^{-\sigma/2}$$

is known. In view of this,

$$z(t) = \int_t^T S_p(s-t) \psi_w(s) ds$$

belongs to $C([0, T], W_p^\sigma(\Omega)) \subseteq C([0, T], C(\bar{\Omega}))$, if $p > (1 - (\sigma/2))^{-1}$. The latter holds true for $p > 3$, as $\sigma < 1 + 1/p$. By $n \geq 2$ and $p > n + 1$ we have $p > 3$. The third part of h is bounded and measurable, too. By $\Phi_i \in W_p^\sigma(\Omega)$ we find as above that $S_q(s-t)\Phi_i = S_p(s-t)\Phi_i$; hence $v_i(t, s) = S_p(s-t)\Phi_i$ belongs to $C(D, W_p^\sigma(\Omega)) \subseteq C(D, C(\bar{\Omega}))$, where $D = \{(t, s) \in [0, T] \times [0, T] \setminus \{(t, s) | 0 \leq t \leq s\}\}$. Therefore the abstract Riemann–Stieltjes integrals

$$\int_t^T \tau S_q(s-t) \Phi_i d\lambda_i(s) \quad (i = 1, \dots, k)$$

exist and belong to the class of abstract functions of bounded variation on $[0, T]$ with values in $C(\Gamma)$. \square

LEMMA 4.3. *Assumption (A2) is satisfied under the assumptions imposed on Φ_1, \dots, Φ_n .*

Proof. We begin with (ii). All entries of $\chi''(t, x)$, $b''(t, x)$, and the functions φ_{ww} , ψ_{ww} , $p(t, x)$ are bounded and measurable. Therefore by (4.6),

$$\begin{aligned} |\mathcal{L}''(v^\circ, u^\circ)[h_1, h_2]| &\leq c_1 \|h_1\|_2 \|h_2\|_2 \\ &\quad + c_2 \|LB'h_1\|_{L_2(0, T; L_2(\Omega))} \|LB'h_2\|_{L_2(0, T; L_2(\Omega))} \\ &\quad + c_3 \|\Lambda B'h_1\|_{L_2(\Omega)} \|\Lambda B'h_2\|_{L_2(\Omega)}. \end{aligned} \tag{4.9}$$

Arguing as above we find

$$(Lz)(t) = \int_0^t A_2 S_2(t-s) N_2 z(s) ds,$$

$$\Lambda z = \int_0^T A_2 S_2(T-s) N_2 z(s) ds$$

for $z \in X_p$. The operators on the right-hand side are continuous from X_2 to $L_2(0, T; L_2(\Omega))$ and $L_2(\Omega)$, respectively (we use (2.6) for $r = 2$ and results about weakly singular integral operators in Krasnosel'skij [14]). B maps continuously $X_2 \times X_2$ into X_2 . Now (A2)(ii), follows directly from (4.9).

(i) By means of the second-order Taylor expansion of $\beta(t) = \mathcal{L}((v^o, u^o) + th, y)$ at $t = 0$,

$$2r_2^{\mathcal{L}}(h) = \int_0^T \int_{\Gamma} h(t, x)^T \{[\chi''_{\nu}(t, x) - \chi''(t, x)] + p(t, x)[b''_{\nu}(t, x) - b''(t, x)]\} h(t, x) dS_x dt$$

$$+ \int_0^T \int_{\Omega} [\psi_{ww}^{\nu} - \psi_{ww}](t, x) ((LB'h)(t, x))^2 dx dt$$

$$+ \int_{\Omega} [\varphi_{ww}^{\nu} - \varphi_{ww}](x) ((\Lambda B'h)(x))^2 dx,$$

where $\nu \in (0, 1)$ is independent of (t, x) , $h = (v, u)$,

$$\psi_{ww}^{\nu}(t, x) = \psi_{ww}(t, x, w^o(t, x) + \nu w(t, x))$$

($w = Lv$), and χ''_{ν} , b''_{ν} , φ_{ww}^{ν} are defined analogously at $(v^o + \nu v, u^o + \nu u)$ and $w^o(T) + \nu w(T)$, respectively. All terms in brackets tend to zero in L_{∞} as $\|h\|_{\infty} \rightarrow 0$ owing to the continuity of the corresponding Nemytskij operator in U_{∞} . The other parts can be estimated by $c\|h\|_2^2$. This yields (A2)(ii). \square

LEMMA 4.4. For all $z \in L_p(0, T; L_p(\Gamma))$,

$$\|(I - KB_v)^{-1} z\|_2 \leq c\|z\|_2.$$

Proof. It is well known that for $z \in X_p$ the Bochner integral equation

$$x(t) - \int_0^t \tau A S(t-s) N b_w(s) x(s) ds = z(t)$$

admits a unique solution $x \in X_p$. Arguing as in the proof of Lemma 4.2 we have

$$x(t) - \int_0^t \tau A_2 S_2(t-s) N_2 b_w(s) x(s) ds = z(t),$$

hence (with a generic c),

$$\begin{aligned}\|x(t)\|_{L_2(\Gamma)} &\leq c \int_0^t \|\tau A_2 S_2(t-s) N_2\|_{L_2(\Gamma) \rightarrow L_2(\Gamma)} \|x(s)\|_{L_2(\Gamma)} ds + \|z(t)\|_{L_2(\Gamma)} \\ &\leq \|z(t)\|_{L_2(\Gamma)} + c \int_0^t (t-s)^{-\frac{1}{4}+\varepsilon} \|x(s)\|_{L_2(\Gamma)} ds\end{aligned}$$

by (2.6), where $\varepsilon > 0$ can be taken arbitrarily small. This is a weakly singular integral inequality with positive kernel. Therefore, $\|x(t)\|$ can be estimated by $\|x(t)\| \leq \alpha(t)$, where

$$\alpha(t) = \|z(t)\|_{L_2(\Gamma)} + c \int_0^t (t-s)^{-\frac{1}{4}+\varepsilon} \alpha(s) ds.$$

(cf. Dixon and McKee [5]).

Now it follows from the theory of weakly singular integral equations for real functions that

$$\|\alpha\|_{L_2(0,T)} \leq c \left(\int_0^T \|z(t)\|_{L_2(\Gamma)}^2 dt \right)^{1/2} = c \|z\|_2;$$

hence $\|x\|_2 \leq c \|z\|_2$, too. \square

LEMMA 4.5. *Let (v^o, u^o) be regular and admissible. Then assumption (A1) is satisfied for problem (P).*

Proof. Let (v, u) be an admissible pair for problem (P). Then

$$(4.10) \quad v = KB(v, u) + \tau d,$$

$$(4.11) \quad G_i(LB(v, u) + d)(t) \leq c_i(t) \quad (i = 1, \dots, k).$$

From the Taylor expansion of B at (v^o, u^o) we obtain

$$(4.12) \quad v - v^o = K[B_v(v - v^o) + B_u(u - u^o)] + Kr_1^B(v - v^o, u - u^o)$$

and

$$(4.13) \quad G_i(w^o) + G_i L[B_v(v - v^o) + B_u(u - u^o)] + G_i Lr_1^B(v - v^o, u - u^o) \leq c_i$$

($i = 1, \dots, k$) (note that $w^o = LB(v^o, u^o) + d$).

Now consider the pair (v_1, u) solving the linearized state equation

$$(4.14) \quad v_1 - v^o = K[(B_v(v_1 - v^o) + B_u(u - u^o))].$$

Subtracting (4.14) from (4.12),

$$(4.15) \quad v - v_1 = KB_v(v - v_1) + Kr_1^B(v - v^o, u - u^o).$$

By Lemma 4.4,

$$(4.16) \quad \|v - v_1\|_2 \leq c \|Kr_1^B(v - v^o, u - u^o)\|_2 \leq c \|r_1^B(v - v^o, u - u^o)\|_2$$

(with generic c). It is known that for the Nemytskij operator B

$$(4.17) \quad \|r_1^B(v - v^o, u - u^o)\|_2 \|(v - v^o, u - u^o)\|_2^{-1} \rightarrow 0$$

as $\|(v - v^o, u - u^o)\|_\infty \rightarrow 0$. Therefore $k_1 = v_1 - v^o$, $z_1 = u - u^o$ could be a candidate for (A1), but (k_1, z_1) will possibly not fulfil the linearized constraints. (4.13) yields

$$(4.18) \quad G_i(w^o) + G_i L(B_v k_1 + B_u z_1) \leq c_i - G_i L(B_v(v - v_1) + r_1^B(v - v^o, u - u^o)),$$

$i = 1, \dots, k$. From the second assertion of Lemma 3.2 we conclude that $G_i L$ is continuous from X_2 to $C[0, T]$ (note that $(AS(t-s)N)^* \Phi_i = \tau S_q \Phi_i = \tau S_p \Phi_i$). Hence

$$\max_{[0, T]} |G_i L(B_v(v - v_1) + r_1^B)(t)| \leq \alpha_i(c + 1) \|r_1^B\|_2 \leq c^1 \|r_1^B\|_2$$

$i = 1, \dots, k$. As (v^o, u^o) is regular, there are (\bar{v}, \bar{u}) and a $\delta > 0$ such that

$$(4.19) \quad \bar{v} - v^o = K(B_v(\bar{v} - v^o) + B_u(\bar{u} - u^o)),$$

$$(4.20) \quad (G_i(w^o) + G_i L(B_v(\bar{v} - v^o) + B_u(\bar{u} - u^o)))(t) \leq c_i(t) - \delta$$

$i = 1, \dots, k$, $t \in [0, T]$. We put $\varepsilon = c^1 \|r_1^B\|_2$, $\lambda = \varepsilon/(\varepsilon + \delta)$,

$$u_2 = (1 - \lambda)u + \lambda \bar{u}, \quad v_2 = (1 - \lambda)v_1 + \lambda \bar{v}.$$

Then the pair $(v_2 - v^o, u_2 - u^o)$ belongs to $M(v^o, u^o)$. This follows simply from a convex combination of (4.14), (4.19) and (4.18), (4.20), respectively. We take $k = v_2 - v^o$, $z = u_2 - u^o$ and find

$$\begin{aligned} \|(k, z) - (v - v^o, u - u^o)\|_2 &\leq \|v_2 - v\|_2 + \|u_2 - u\|_2 \\ &\leq \|v - v_1\|_2 + \|v_2 - v_1\|_2 + \lambda \|\bar{u} - u\|_2 \\ &\leq c \|r_1^B\|_2 + \lambda (\|\bar{v} - v_1\|_2 + \|\bar{u} - u\|_2) \\ &\leq \|r_1^B\|_2 (c + \delta^{-1} c^1 (\|\bar{v} - v_1\|_2 + \|\bar{u} - u\|_2)), \end{aligned}$$

by (4.16) and the definition of λ . For $\|(v - v^o, u - u^o)\|_\infty \rightarrow 0$ we have $v_1 \rightarrow v^o$; hence the term in the last bracket remains bounded. The proof is completed by (4.17). \square

Summarizing, Theorem 4.1 and Lemmas 4.2–4.5 yield the main theorem.

THEOREM 4.6. *Let (v^o, u^o) be regular and admissible for the optimal control problem (P), $w^o = d + LB(v^o, u^o)$.*

If (v^o, u^o) satisfies the second-order condition (4.2), then (v^o, u^o) is locally optimal for (P), and (4.3) holds.

COROLLARY 1. *Under the assumptions of Theorem 3, there are $\rho > 0$, $\alpha > 0$, such that (4.3) holds for all admissible (v, u) such that $\|u - u^o\|_\infty \leq \rho$. Thus u^o is a locally optimal control.*

(The corollary follows from (4.3), since $\|v - v^o\|_\infty \rightarrow 0$ for $\|u - u^o\|_\infty \rightarrow 0$.)

COROLLARY 2. *If b and χ satisfy conditions (4.4) and (4.5), respectively, then Theorem 3 and Corollary 1 remain valid for $U_\infty := L_p(0, T; L_p(\Gamma))$ and $\|u - u^o\|_\infty := \|u - u^o\|_p$.*

This is a simple consequence of Remark 2.

Remark 3.

1. The method of this paper extends also to more general optimal control problems with additional control distributed in Ω . The equation of state would be of the type

$$\begin{aligned}w_t(t, x) &= (\Delta - 1)w(t, x) + b_1(t, x, w(t, x), u_1(t, x)) \\w(0, x) &= w_0(x) \\ \frac{\partial w}{\partial n}(t, x) &= b_2(t, x, w(t, x), u_2(t, x)).\end{aligned}$$

Introducing $v(t, x) = \tau w(t, x)$ this leads to a system of Bochner integral equations for the state $(v(t, x), w(t, x))$. However, the presentation of the theory is notationally much more complex.

2. State constraints of the form

$$\int_{\Omega} \langle \Phi_i(x), \nabla_x w(t, x) \rangle dx \leq c_i(t)$$

can be transformed to (2.3) integrating by parts, provided that

$$\Phi_j \in (W_p^{1+\sigma}(\Omega))^n \cap (\overset{\circ}{W}_p^1(\Omega))^n .$$

5. Verification of the second-order condition. To verify the strict positivity of quadratic forms is a difficult task in general. This refers also to (4.2). It is well known from the optimal control theory for systems of ordinary differential equations that matrix Riccati equations may be helpful to solve this problem, see, for instance, Bryson and Ho [4] or Malanowski [19]. A similar approach works for parabolic equations, where the control is distributed, i.e., acting only within the domain under consideration. In this way, parabolic equations of Riccati type are obtained for the kernels representing certain operator-valued functions. We refer to Lions [18, Chap. 3]. This method cannot be extended directly to boundary control problems.

On the other hand, even the solution of parabolic Riccati equations is a difficult question, which generally can only be answered numerically. Therefore, we propose the reduction of the problem to one for a system of ordinary differential equations by means of a finite element method.

We have

$$\mathcal{L}''((v^o, u^o))[(k, z), (k, z)] = Q(w, z),$$

where

$$\begin{aligned}Q(w, z) &= \int_0^T \int_{\Gamma} (w(t, x), z(t, x)) Q_1(t, x) (w(t, x), z(t, x))^T dS_x dt \\ &\quad + \int_0^T \int_{\Omega} Q_2(t, x) w(t, x)^2 dx dt \\ &\quad + \int_{\Omega} \Phi_{ww}(x) w(T, x)^2 dx, \\ &= Q_{11}(w, w) + Q_{12}(w, z) + Q_{22}(z, z),\end{aligned}$$

Q_1 is a certain 2×2 matrix with L_∞ entries, $Q_2 \in L_\infty$ (cf. (4.6)) and w solves

$$(5.1) \quad \begin{aligned} w_t(t, x) &= (\Delta - 1)w(t, x), \\ w(0, x) &= 0, \\ \frac{\partial w}{\partial n}(t, x) &= b_w(t, x)w(t, x) + b_u(t, x)z(t, x). \end{aligned}$$

Let $V_h \subset H^1(\Omega)$ be a finite element space depending on a discretization parameter $h > 0$, $V_h = \text{span}\{v_1, \dots, v_m\}$. We approximate (5.1) by the finite element scheme

$$(5.2) \quad \begin{aligned} \int_{\Omega} \left[\frac{d}{dt} w_h(t, x) v(x) + w_h(t, x) v(x) + \nabla w_h(t, x) \cdot \nabla v(x) \right] dx \\ = \int_{\Gamma} (b_w(t, x) w_h(t, x) + b_u(t, x) z(t, x)) v(x) dS_x, \\ w_h(0) = 0 \end{aligned}$$

for all $v \in V_h$, where $w_h(t, x) = \sum_{i=1}^m w_i(t) v_i(x)$. It can be shown that under natural assumptions on V_h

$$(5.3) \quad \begin{aligned} \max_{t \in [0, T]} \|w(t, \cdot) - w_h(t, \cdot)\|_{L_2(\Omega)}^2 + \int_0^T \|w(t, \cdot) - w_h(t, \cdot)\|_{H^1(\Omega)}^2 dt \\ \leq ch^\alpha \int_0^T \|z(t)\|_{L_2(\Gamma)}^2 dt \end{aligned}$$

where $\alpha > 0$. The proof is based on a technique, developed by Lasiecka [15] for boundary value problems with L_2 boundary data.

Equation (5.2) is equivalent to a system of ordinary differential equations for the vector valued function $w(t) = (w_1(t), \dots, w_m(t))$.

THEOREM 5.1. *Suppose that the error estimate (5.3) holds true. Then*

$$(5.4) \quad Q(w, z) \geq \delta \|z\|_2^2$$

if and only if

$$(5.5) \quad Q(w_h, z) \geq (\delta - \varepsilon) \|z\|_2^2$$

for all $\varepsilon > 0$ and all $h \leq h_0(\varepsilon)$.

Proof. Simple estimations yield

$$\begin{aligned} |Q_{11}(w_1, w_2)| &\leq c_1(\|w_1\|_C \|w_2\|_C + \|w_1\|_{H^1} \|w_2\|_{H^1}), \\ |Q_{12}(w, z)| &\leq c_2 \|w\|_{H^1} \|z\|_2, \end{aligned}$$

where $\|w\|_C = \|w\|_{C([0, T], L_2(\Omega))}$, $\|w\|_{H^1} = \|w\|_{L_2(0, T; H^1(\Omega))}$. Let (5.4) be satisfied. Then

$$(5.6) \quad \begin{aligned} Q(w_h, z) &= Q(w, z) + Q_{11}(w + w_h, w_h - w) + Q_{12}(w_h - w, z) \\ &\geq \delta \|z\|_2^2 - c_1(\|w + w_h\|_C \|w_h - w\|_C + \|w + w_h\|_{H^1} \|w_h - w\|_{H^1}) \\ &\quad - c_2 \|w_h - w\|_{H^1} \|z\|_2 \\ &\geq (\delta - ch^{\alpha/2}) \|z\|_2^2, \end{aligned}$$

as

$$\max\{\|w\|, \|w_h\|\} \leq c\|z\|_2$$

(see [28]) and (5.3) is true. Equation (5.5) is a consequence of (5.6).

Conversely, if (5.5) holds, then as in (5.6),

$$Q(w, z) \geq (\delta - \varepsilon)\|z\|_2^2 - ch^{\alpha/2}\|z\|_2^2$$

for all $h \leq h_0(\varepsilon)$. Equation (5.4) follows from $h \downarrow 0$, as ε was arbitrary. Theorem 5.1 permits us to investigate the positivity of the quadratic form Q for all solutions of a system of ordinary differential equations, where the known theory of Riccati equations applies. We will not discuss the details further.

Acknowledgment. The authors are very grateful to Prof. N. Weck for a valuable remark improving Lemma 4.2.

REFERENCES

- [1] W. ALT AND U. MACKENROTH, *On the numerical solution of state constrained coercive parabolic optimal control problems*, Optimal Control Partial Diff. Equations, Internat. Schriftenreihe Numer. Math., 68 (1984), pp. 44–62.
- [2] H. AMANN, *Parabolic evolution equations with nonlinear boundary conditions*, in Proc. Sympos. Pure Math., Vol. 45, Part I, Nonlinear Functional Analysis, F. E. Browder, ed., 1986, pp. 17–27.
- [3] ———, *Parabolic evolution equations with nonlinear boundary conditions*, J. Differential Equations, 72 (1988), pp. 201–269.
- [4] A. E. BRYSON AND Y. C. HO, *Applied optimal control*, Blaisdell, Waltham, MA, 1969.
- [5] J. DIXON AND S. MCKEE, *Weakly singular discrete Gronwall inequalities*, Z. Angew. Math. Mech., 66 (1986), pp. 535–544.
- [6] H. O. FATTORINI, *Optimal control problems for semilinear parabolic distributed parameter systems*, to appear.
- [7] ———, *Boundary control systems*, SIAM J. Control Optim., 6 (1968), pp. 349–385.
- [8] ———, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Appl. Math. Optim., 15 (1987), pp. 141–185.
- [9] A. FRIEDMAN, *Optimal control for parabolic equations*, J. Math. Anal. Appl., 18 (1967), pp. 479–491.
- [10] H. GOLDBERG AND F. TRÖLTZSCH, *Second order optimality conditions for a class of control problems governed by nonlinear integral equations with application to parabolic boundary control*, Optimization, 20 (1989), pp. 687–698.
- [11] ———, *Second order optimality conditions for nonlinear parabolic boundary control problems*, in Proc. Conf. on Optimal Control of Partial Differential Equations, Irsee, 1990; Lecture Notes Control Inform. Sci., Vol. 149, K. H. Hoffmann and W. Krabs, eds., 1991, pp. 93–103.
- [12] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum 3: Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.
- [13] G. KNOWLES, *Finite element approximation of parabolic time optimal control problems*, SIAM J. Control Optim., 20 (1982), pp. 414–427.
- [14] M. A. KRASNOSEL'SKIJ, P. P. ZABREJKO, E. I. PUSTYL'NIK, AND P. E. SOBOLEVSKIJ, *Linear operators in spaces of summable functions*, Nauka, Moscow, 1966. (In Russian.)
- [15] I. LASIECKA, *Boundary control of parabolic systems: finite-element approximation*, Appl. Math. Opt., 6 (1980), pp. 31–62.
- [16] ———, *Unified theory for abstract parabolic boundary problems—a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–333.
- [17] ———, *Ritz-Galerkin approximation of abstract parabolic boundary value problems with rough boundary data— L_p -theory*, Math. Comp., 47 (1986), pp. 55–75.
- [18] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunvol Gauthier-Villars, Paris, 1968.
- [19] R. MALANOWSKI, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.

- [20] H. MAURER, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 163–177.
- [21] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Springer Verlag, New York, 1983.
- [22] E. SACHS, *A parabolic control problem with a boundary condition of the Stefan–Boltzman type*, Z. Angew. Math. Mech., 58 (1978), pp. 443–449.
- [23] E. J. P. G. SCHMIDT, *Boundary control for the heat equation with non-linear boundary condition*, J. Differential Equations, 78 (1989), pp. 89–121.
- [24] F. TRÖLTZSCH, *Optimality conditions for parabolic control problems and applications*, Teubner–Texte zur Mathematik, 62, B.G. Teubner Verlagsgesellschaft, Leipzig, 1984.
- [25] ———, *On changing the spaces in Lagrange multiplier rules for the optimal control of non-linear operator equations*, Optimization, 16 (1985), pp. 877–885.
- [26] ———, *On the semigroup approach for the optimal control of semilinear parabolic equations including distributed and boundary control*, Z. Anal. Anwendungen, 8 (1989), pp. 431–443.
- [27] ———, *Approximation of non-linear parabolic boundary control problem by the Fourier method-convergence of optimal controls*, Optimization, 22 (1991), pp. 83–98.
- [28] ———, *On convergence of semidiscrete Ritz–Galerkin schemes applied to the boundary control of parabolic equations with non-linear boundary condition*, Z. Angew. Math. Mech., 72 (1992), pp. 291–301.
- [29] L. VON WOLFERSDORF, *Optimal control for processes governed by mildly nonlinear differential equations of parabolic type*, Z. Angew. Math. Mech., 56/77 (1976/1977), pp. 531–538/11–17.

A LINEAR ALGEBRAIC FRAMEWORK FOR THE ANALYSIS OF DISCRETE-TIME NONLINEAR SYSTEMS*

J. W. GRIZZLE†

Abstract. A linear algebraic framework for the analysis of synthesis-type problems for discrete-time nonlinear systems is introduced. This is an extension of a similar tool for continuous-time systems that established important connections between many algorithms associated with right-invertibility, left-invertibility and dynamic decoupling, as well as between these algorithms and an approach based upon differential algebra. A similar payoff is seen to be possible in the discrete-time setting.

Key words. nonlinear systems, discrete-time, ranks, invertibility

AMS subject classifications. 93C10, 49E05

1. Introduction. This paper extends to the class of discrete-time nonlinear systems the linear algebraic framework of [4], which has proven useful in the analysis of several synthesis problems for the class of continuous-time nonlinear systems [1]–[3], [13], [16], [32]. Recall that [4], through the introduction of a chain of subspaces naturally associated with the output of a system, provided a high-level interpretation of the inversion and dynamic decoupling algorithms that are built around the recursive computation of certain ranks associated with left-invertibility, right-invertibility, and noninteracting control. In addition, it established relationships between these algorithms and the differential algebraic approach. This same linear algebraic setting has been used in [3] to formulate in an intrinsic way the regularity (constant rank) conditions common to several procedures for synthesizing nonlinear dynamic compensators.

The reader is reminded that the importance of algebraic techniques and reasoning for analyzing many aspects of discrete-time nonlinear systems has been firmly established in [8]–[12], [25], [28], [29], and the references therein. The techniques employed here are most closely related to those of [10] and [28].

When studying continuous-time nonlinear systems, the class of affine systems (so-called because the dynamics is affinely parametrized by the control variables) has received the bulk of the attention of the nonlinear community. This is true for several reasons, the most important of which is that the class of affine systems is general enough to encompass many models arising in practice. However, it is also specific enough to admit reasonably simple analyses from at least two perspectives: *geometrically*, we are working with a finite number of vector fields, a drift term, and m control vector fields, as opposed to some arbitrarily, smoothly parametrized family of vector fields; *algebraically*, the various derivatives of the outputs depend polynomially on the inputs and their derivatives (with coefficients depending on the state, and the highest order derivative of the input appearing affinely), as opposed to some more general nonlinear dependence on the input. However, if we accept as an axiom that

* Received by the editors May 13, 1991; accepted for publication (in revised form) March 23, 1992. This work was completed while the author was a visiting professor (Poste Rouge, Centre National de la Recherche Scientifique) at the Laboratoire des Signaux et Systèmes, École Supérieure d'Électricité-Centre National de la Recherche Scientifique, Gif-sur-Yvette, France. This research was supported by the National Science Foundation contract NSF ECS-88-96136.

† Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109-2122.

any interesting class of discrete-time systems should include time-sampled (digital) versions of the class of continuous-time affine systems, then we are obliged to consider systems of the form

$$\begin{aligned}x[k+1] &= f(x[k], u[k]), \\ y[k] &= h(x[k], u[k]),\end{aligned}$$

where f and h are sufficiently smooth functions, but otherwise arbitrary (consider sampling a continuous-time bilinear system). Consequently, it is not possible to assume that the dynamics is affine in the control variable (and hence, finitely parametrized); and even if it were, this would not entail that the iterates of the output depend polynomially on the inputs, with the highest-order delayed input appearing affinely. Consequently, the proof techniques of [4], based upon “global” interpretations of the inversion and dynamic extension algorithms, cannot be easily extended to discrete-time systems; a more intrinsic, “algorithm-free” analysis will be performed.

In §2 of this paper, the linear algebraic framework of [4] is developed for analytic discrete-time systems, thereby extending the notion of the rank of a system introduced in [10]. This includes the definition of a chain of subspaces constructed from the outputs of the system and their iterates and an analysis of the convergence properties of the chain of subspaces. It is noted that when the function f describing the dynamics is not a submersion, certain new phenomena can occur, requiring a slightly different analysis involving a combination of geometric and algebraic reasoning. Section 3 collects a few results that are useful for establishing relations between some existing work involving rank computations in an “algorithmic” form and the linear algebraic setting proposed here. Section 4 relates the abstract notion of rank, introduced in §2, to the injectivity and surjectivity properties of certain maps strongly connected with left- and right-invertibility. Finally, §5 points out the links between the approach used in this paper and that of [10]; §6 shows the affinity with the work of [28].

2. Rank and structure at infinity. The notion of the *rank* of a nonlinear system was introduced by Fliess in [8] and yielded fundamental results on right- and left-invertibility and noninteracting control of continuous-time systems. Extensions to a class of rational discrete-time systems have been given in [10], using difference algebra in place of differential algebra. Here, using elementary vector space techniques as in [4], the rank of a discrete-time system will be generalized to analytic systems admitting a global state space representation on \mathbb{R}^n . This may be a strong assumption.

2.1. Linear algebraic framework. Consider a discrete-time system

$$(2.1) \quad \Sigma: \quad \begin{aligned}x[k+1] &= f(x[k], u[k]), \\ y[k] &= h(x[k], u[k]),\end{aligned}$$

where $x[k] \in X = \mathbb{R}^n$, $u[k] \in U = \mathbb{R}^m$, $y[k] \in Y = \mathbb{R}^\mu$, f and h are analytic functions of their arguments, and $x[0] = x_0$. It is convenient to let $f^u(x) := f(x, u)$ so that we may write:

$$\begin{aligned}x[1] &= f^{u[0]}(x_0), \\ x[2] &= f^{u[1]}(x[1]) = f^{u[1]} \circ f^{u[0]}(x_0) \\ &\vdots \\ x[k] &= f^{u[k-1]} \circ \dots \circ f^{u[0]}(x_0),\end{aligned}$$

where \circ denotes composition. Then, since $y[k] = h^{u[k]}(x[k])$, where $h^u(x) := h(x, u)$,

$$(2.2) \quad \begin{aligned} y[k] &= y[k](x_0, u[0], \dots, u[k]) \\ &= h^{u[k]} \circ f^{u[k-1]} \circ \dots \circ f^{u[0]}(x_0). \end{aligned}$$

Because (2.1) is time-invariant,

$$(2.3) \quad \begin{aligned} y[k+1] &= y[k](x[1], u[1], \dots, u[k+1]) \\ &= y[k](f(x_0, u[0]), u[1], \dots, u[k+1]), \end{aligned}$$

which will be important later when establishing a certain finiteness property.

Let \mathcal{R}_k denote the ring of real analytic functions of the components of $(x, u[0], \dots, u[k])$, and let \mathcal{K}_k be the associated field of fractions, that is, the field of *meromorphic* functions in the variables $(x, u[0], \dots, u[k])$. A typical element of \mathcal{K}_k would have the form $\eta(v) = \pi(v)/\theta(v)$, where π and θ are elements of \mathcal{R}_k , θ is not the zero function, and $v = (v_1, \dots, v_j)$ denotes the various components of $(x, u[0], \dots, u[k])$. Recall that $\partial/\partial v_i$ acting on η is formally defined by the usual quotient rule of calculus,

$$(2.4) \quad \frac{\partial}{\partial v_i} \frac{\pi(v)}{\theta(v)} := \left(\theta(v) \frac{\partial}{\partial v_i} \pi(v) - \pi(v) \frac{\partial}{\partial v_i} \theta(v) \right) / \theta^2(v),$$

and the formal differential of η is

$$(2.5) \quad d\eta(v) := \sum_{i=1}^j \frac{\partial \eta(v)}{\partial v_i} dv_i.$$

Let \mathcal{E} denote the vector space over $\mathcal{K} := \mathcal{K}_n$ spanned by $\{dx_1, \dots, dx_n, du_1[0], \dots, du_m[0], du_1[n], \dots, du_m[n]\}$. Note that \mathcal{E} is a finite-dimensional vector space; indeed, its dimension is $n + (n+1)m$. For notational convenience, $\{dx_1, \dots, dx_n\}$ will simply be written as dx , $\{du_1[0], \dots, du_m[0]\}$ as $du[0]$, etc., so that $\mathcal{E} = \text{span}_{\mathcal{K}}\{dx, du[0], \dots, du[n]\}$.

Observe now, for all $0 \leq k \leq n$ and $1 \leq j \leq \mu$, that $dy_j[k] \in \mathcal{E}$, since

$$(2.6) \quad dy_j[k] = \sum_{i=1}^n \frac{\partial y_j[k]}{\partial x_i} dx_i + \sum_{\ell=0}^k \sum_{i=1}^m \frac{\partial y_j[k]}{\partial u_i[\ell]} du_i[\ell].$$

Define a chain of subspaces $\mathcal{E}_0 \subset \dots \subset \mathcal{E}_n$ of \mathcal{E} by [4] (see also [7])

$$(2.7) \quad \mathcal{E}_k := \text{span}_{\mathcal{K}}\{dx, dy[0], \dots, dy[k]\}$$

and the associated list of dimensions $\rho_0 \leq \rho_1 \leq \dots \leq \rho_n$ by

$$(2.8) \quad \rho_k = \dim_{\mathcal{K}} \mathcal{E}_k.$$

We emphasize that $dy[k]$ denotes $\{dy_1[k], \dots, dy_{\mu}[k]\}$ and that this abuse of notation will be used quite often to keep the notation compact.

It will turn out for *generically submersive* systems,¹ that is, for systems where

$$(2.9) \quad n = \text{rank}_{\mathcal{K}} \left[\frac{\partial f}{\partial x}(x, u[0]) : \frac{\partial f}{\partial u}(x, u[0]) \right],$$

¹ The mathematical importance of this assumption will be seen in the next subsection; in terms of control systems, it means that with a feedback, the drift dynamics could be made (generically) invertible, creating a kind of group action.

that

$$(2.10) \quad \rho^* := \rho_n - \rho_{n-1}$$

is a limiting value of the chain (2.7) in the sense that if we were to extend the chain in the obvious manner, then $\rho_{n+r} = \rho_n + r\rho^*$, for all integers $r \geq 0$. Many system models of the form (2.1) would satisfy (2.9), since it is equivalent to $f(\mathbb{R}^n, \mathbb{R}^m)$ having nonempty interior in \mathbb{R}^n , and this is a *necessary* condition for accessibility [17]. It is always satisfied for a time-sampled representation of a continuous-time system. Moreover, it has just been established in [11] that, in a certain sense, rational input-output systems admit *local* state space representations satisfying condition (2.9). To avoid passing to a local representation, the following construction is used here in the general case where (2.9) is not satisfied.

Let $\mathcal{K}^+ := \mathcal{K}_{2n}$, and define $\mathcal{E}^+ := \text{span}_{\mathcal{K}^+}\{dx, du[0], \dots, du[2n]\}$. Define a chain of subspaces $\mathcal{E}_0^+ \subset \dots \subset \mathcal{E}_n^+$ of \mathcal{E}^+ by

$$(2.11) \quad \mathcal{E}_k^+ := \text{span}_{\mathcal{K}^+}\{dx, du[0], \dots, du[n-1], dy[n], \dots, dy[n+k]\}$$

and the associated list of dimensions $\rho_0^+ \leq \dots \leq \rho_n^+$ by

$$(2.12) \quad \rho_k^+ := \dim \mathcal{E}_k^+.$$

Then, even without condition (2.9), it will turn out that

$$(2.13) \quad \rho^{+*} := \rho_n^+ - \rho_{n-1}^+$$

is a limiting value in the sense discussed earlier for ρ^* . Whenever the system (2.1) is generically submersive, it will be established that

$$(2.14) \quad \rho_k^+ = \rho_k + nm, \quad k \geq 0,$$

so that $\rho^{+*} = \rho^*$.

Anticipating these technical results, ρ^{+*} is defined to be the *rank* [10] of the system (2.1).

Remarks. (a) In [6], it is shown (for continuous-time systems) that the chain of subspaces (2.7) is closely related to classical objects in algebra, namely filtrations, and consequently, Hilbert polynomials; a similar result is true in discrete-time [7]. One of the main points of the analysis presented in this paper is the establishment of a priori bounds on the number of steps required to compute the limiting ranks of the filtrations whenever the system has a standard state-space representation; such bounds are not provided by the classical results of algebra, which, on the other hand, apply to more general situations.

(b) In analogy with [4], [22], the list of integers $\{\sigma_0, \dots, \sigma_n\}$ defined by

$$(2.15) \quad \sigma_i = \rho_i - \rho_{i-1}, \quad 0 \leq i \leq n$$

with the convention that $\rho_{-1} = n$, could be called the *transient structure at infinity*, while the list of integers $\{\sigma_0^+, \dots, \sigma_n^+\}$, defined by

$$(2.16) \quad \sigma_i^+ := \rho_i^+ - \rho_{i-1}^+, \quad 0 \leq i \leq n,$$

with the convention that $\rho_{-1}^+ = n + nm$ could be called the *persistent structure at infinity*. For generically submersive systems the two lists coincide, and we can speak simply of the *structure at infinity*; this is also the case for systems satisfying certain constant rank hypotheses in the neighborhood of an equilibrium point, as can be seen from the results of [18], [24], including (constant-coefficient) linear systems.

2.2. Convergence of the chain $\mathcal{E}_0 \subset \cdots \subset \mathcal{E}_n$. The goal of this subsection is to justify terminating the chain (2.7) at n , the dimension of the state space of (2.1), whenever the system is generically submersive. For a linear system, this would follow as an easy consequence of the Cayley–Hamilton theorem; in the case of nonlinear systems, more work is required.

Let $k \geq 0$ be any nonnegative finite integer, and recall that \mathcal{K}_k is the field of meromorphic functions of $(x, u[0], \dots, u[k])$. Before, when defining \mathcal{E}_k ($0 \leq k \leq n$), the span was taken with respect to $\mathcal{K} := \mathcal{K}_n$. It is easily seen that the dimension of \mathcal{E}_k does not change if instead the span is taken with respect to \mathcal{K}_k . For $k > n$, define \mathcal{E}_k in the obvious way, following (2.7), taking the span with respect to \mathcal{K}_k .

THEOREM 2.1. *Suppose that (2.1) is generically submersive. Then, for all integers $k \geq n$, $\dim \mathcal{E}_k - \dim \mathcal{E}_{k-1} = \dim \mathcal{E}_n - \dim \mathcal{E}_{n-1}$; that is, $\rho_k - \rho_{k-1} = \rho_n - \rho_{n-1}$.*

The proof of the theorem will be divided into several parts, each establishing a particular property of the chain (2.7) arising from the recursive manner in which the functions $y[k]$ are constructed from the system. Define $\delta : \mathcal{K}_{k-1} \rightarrow \mathcal{K}_k$ by

$$(2.17) \quad (\delta\eta)(x, u[0], \dots, u[k]) = \eta(f(x, u[0]), u[1], \dots, u[k]);$$

it is important that f be generically submersive, for otherwise, $\delta\eta$ may not be a meromorphic function (see (2.20) below). This induces an \mathbb{R} -linear mapping² $\Delta : \text{span}_{\mathcal{K}_{k-1}} \{d\eta | \eta \in \mathcal{K}_{k-1}\} \rightarrow \text{span}_{\mathcal{K}_k} \{d\lambda | \lambda \in \mathcal{K}_k\}$ by

$$(2.18) \quad \begin{aligned} \Delta(d\eta) &:= d(\delta\eta), \\ \Delta(\alpha_1 d\eta_1 + \alpha_2 d\eta_2) &:= \delta(\alpha_1) \Delta(d\eta_1) + \delta(\alpha_2) \Delta(d\eta_2) \end{aligned}$$

for $\eta, \eta_1, \eta_2, \alpha_1, \alpha_2 \in \mathcal{K}_{k-1}$. It should be noted that (2.18) is consistent with the chain rule for differentiation, and that, for instance,

$$(2.19) \quad \begin{aligned} dy_j[k+1] &= d(\delta(y_j[k])) \\ &= \Delta(dy_j[k]) \\ &= \sum_{i=1}^n \delta\left(\frac{\partial y_j[k]}{\partial x_i}\right) dx_i[1] + \sum_{\ell=0}^k \sum_{i=1}^m \delta\left(\frac{\partial y_j[k]}{\partial u_i[\ell]}\right) du_i[\ell+1]. \end{aligned}$$

The following two (equivalent) properties are easily established whenever the system (2.1) is generically submersive:

P1. For all $k \geq 1$, $\dim_{\mathcal{K}_{k-1}} \{dx[k]\} = n$.

P2. For all $k \geq 0$ and $\forall \eta \in \mathcal{R}_k$, if $\eta \neq 0$, then $\delta(\eta) \neq 0$.

As a consequence, if $\eta \in \mathcal{K}_k$, then $\delta(\eta)$ is well defined. To see the peculiarities a nonsubmersive system may exhibit, consider the example

$$(2.20) \quad \begin{aligned} x[k+1] &= 0, \\ y[k] &= x[k]u[k], \end{aligned}$$

where x, y , and u are in \mathbb{R} . The function $1/x$ is meromorphic, but $\delta(1/x)$ is not defined because $\delta(x) \equiv 0$.

The main ingredients of a proof of Theorem 2.1 are now presented. In the following, if S is a set, then $|S|$ denotes its cardinality.

² The fact that Δ is well defined follows easily from $\Delta(0) = 0$.

LEMMA 2.2. *Suppose that (2.1) is generically submersive and that for some $k \geq 0$, $I_0 \subset I_1 \subset \dots \subset I_k \subset \{1, \dots, \mu\}$ are index sets such that $\mathcal{E}_k = \text{span}_{\mathcal{K}_k} \{dx, dy_{i_0}[0], \dots, dy_{i_k}[k] | i_j \in I_j, 0 \leq j \leq k\}$, and $\dim \mathcal{E}_k = n + |I_0| + \dots + |I_{k-1}| + |I_k|$. Then, $\dim \text{span}_{\mathcal{K}_k} \{dx, dy_{i_0}[0], \dots, dy_{i_k}[k], dy_{i_k}[k+1] | i_j \in I_j, 0 \leq j \leq k\} = n + |I_0| + \dots + |I_{k-1}| + 2|I_k|$. In other words, once an output component becomes independent, it remains independent.*

Proof. For the proof, see Appendix A. \square

It immediately follows from the above that there exists a basis of a special form for the chain $\mathcal{E}_0 \subset \mathcal{E}_1 \subset \dots$.

LEMMA 2.3. *Suppose that (2.1) is generically submersive. Then, there exist index sets $I_0 \subset I_1 \subset \dots \subset \{1, \dots, \mu\}$ such that, for all $k \geq 1$, $\{dx, dy_{i_0}[0], \dots, dy_{i_k}[k] | i_j \in I_j, 0 \leq j \leq k\}$ is a basis for \mathcal{E}_k .*

Lemma 2.3 establishes that $\{\sigma_k\} = \{\rho_k - \rho_{k-1}\}$ is a nondecreasing sequence. Since $\sigma_k \leq \min\{m, \mu\}$, it follows that $\{\sigma_k\}$ converges in a finite number of steps. However, this does not allow us to terminate the calculations at the n th step unless the upper bound has already been attained. Considering once again (2.20), we calculate that $(\sigma_0) = 1$, but $(\sigma_1) = 0$; that is, the sequence $\{\sigma_k\}$ of “zeros at infinity of order less than or equal to k ” is not nondecreasing, as is always the case for linear systems and continuous-time nonlinear systems.

LEMMA 2.4. *Suppose that (2.1) is generically submersive, and let I_0, \dots, I_n be as in Lemma 2.3. Then, for each $1 \leq j \leq \mu$, there exists an integer N , $1 \leq N \leq n$, such that*

$$(2.21) \quad dy_j[N] \in \text{span}_{\mathcal{K}_N} \{dy_j[0], \dots, dy_j[N-1], dy_{i_0}[0], \dots, dy_{i_N}[N] | i_k \in I_k, 0 \leq k \leq N\}.$$

Proof. For the proof, see Appendix A. \square

The previous and the following lemmas combine to replace the Cayley–Hamilton theorem, which, in the case of a linear system, proves that the chain (2.7) converges in at most n steps.

LEMMA 2.5. *Suppose that (2.1) is generically submersive, and let I_0, \dots, I_n be as in Lemma 2.3. Suppose that $1 \leq j \leq \mu$, and let N be as in Lemma 2.4. Then, for all $k \geq N$,*

$$(2.22) \quad dy_j[k] \in \text{span}_{\mathcal{K}_k} \{dy_j[0], \dots, dy_j[N-1], dy_{i_0}[0], \dots, dy_{i_k}[k] | \\ \text{for } 1 \leq s \leq N, i_s \in I_s, \text{ and for } s > N, i_s \in I_N\}.$$

Proof. The proof is immediate from (2.18) and Lemma 2.4. \square

The proof of Theorem 2.1 is now given easily. Let $\{I_k\}$ be the collection of index sets determined by Lemma 2.2. By Lemma 2.5, $I_{n+r} = I_n$ for all $r \geq 1$. Hence, the components of the output either become independent by the n th iteration of the dynamics, or they remain dependent for all iterations. Consequently, for all $r \geq 1$, $\dim \mathcal{E}_{n+r} - \dim \mathcal{E}_{n+r-1} = \dim \mathcal{E}_n - \dim \mathcal{E}_{n-1} = \rho^*$.

2.3. Convergence of the chain $\mathcal{E}_0^+ \subset \mathcal{E}_1^+ \subset \dots$. This section addresses the convergence properties of the chain $\mathcal{E}_0^+ \subset \mathcal{E}_1^+ \subset \dots$ for systems that are not necessarily submersive. The idea behind the analysis is that the effect of the nilpotent part of the system on the output sequence is short lived and can be eliminated from the analysis by “ignoring” the first n time instances of the output; this is essentially what

is accomplished by including $\{du[0], \dots, du[n-1]\}$ in the definition of $\mathcal{E}_0^+ \subset \mathcal{E}_1^+ \subset \dots$. The analytical aspects of the proof are based on the following construction.

Let $M_0 := \mathbb{R}^n$ and define for $k \geq 1$

$$(2.23) \quad M_k := x[k] \left(\mathbb{R}^n \times (\mathbb{R}^m)^k \right).$$

Observe that $M_0 \supset M_1 \supset \dots$ since

$$(2.24) \quad M_{k+1} = f(M_k, \mathbb{R}^m).$$

Define $d_0 := n$ and for $k \geq 1$,

$$(2.25) \quad d_k := \text{rank}_{\mathcal{K}_{k-1}} x[k] := \dim \text{span}_{\mathcal{K}_{k-1}} \{dx[k]\}.$$

LEMMA 2.6. *The sequence of integers $\{d_k\}$ is nonincreasing, and if $d_k = d_{k-1}$ then $d_{k+1} = d_k$. Consequently, since d_j can decrease at most n -times, $d_j = d_n$ for all $j \geq n$.*

The proof is given in Appendix B. Very roughly speaking, Lemma 2.6 says that “ $f : M_k \times \mathbb{R}^m \rightarrow M_{k+1}$ behaves like a ‘submersion’ for $k \geq n$ since the ‘dimension’ of M_k is d_k .” Of course, M_k , in general, does not have the structure of a manifold, hence, the imprecision of such a statement. Nevertheless, if we pursue this line of thought for a moment and supposes that $M_0 \supset M_1 \supset \dots$ is a nested sequence of embedded analytic submanifolds, it is clear that, if $\dim M_k = \dim M_{k-1}$, then $\dim M_{k+1} = \dim M_k$, because the condition $\dim M_k = \dim M_{k-1}$ implies that $f : M_{k-1} \times \mathbb{R}^m \rightarrow M_k$ is a submersion. This combined with M_k being open in M_{k-1} gives the result. The proof in Appendix B makes this line of reasoning rigorous with a local analysis, which also establishes the following result: For $k \geq 2$, $j \geq 0$ define $\delta^k : \mathcal{R}_j \rightarrow \mathcal{R}_{j+k}$ by, if $\alpha \in \mathcal{R}_j$, and $0 \leq i \leq k-1$

$$(2.26) \quad \delta^{i+1}(\alpha) = \delta(\delta^i(\alpha)).$$

LEMMA 2.7. *Let $k \geq 0$, $\alpha \in \mathcal{R}_k$. If $\delta^n(\alpha) \neq 0$ (i.e., is not the identically zero function), then $\delta^{n+j}(\alpha) \neq 0$ for all $j \geq 1$.*

Let $\mathcal{R}_k^{\delta^n}$ be the set of real analytic functions

$$(2.27) \quad \mathcal{R}_k^{\delta^n} := \{\delta^n(\alpha) | \alpha \in \mathcal{R}_k\}.$$

$\mathcal{R}_k^{\delta^n} \subset \mathcal{R}_{n+k}$ as a subring, and thus, the associated set of fractions, denoted $\mathcal{K}_k^{\delta^n}$, is a field; indeed, it is a subfield of \mathcal{K}_{n+k} . Note that $\mathcal{K}_k^{\delta^{n+1}} \subset \mathcal{K}_{k+1}^{\delta^n}$.

By Lemma 2.7, for each $k \geq 0$, the two mappings

$$(2.28) \quad \delta : \mathcal{K}_k^{\delta^n} \rightarrow \mathcal{K}_k^{\delta^{n+1}} \subset \mathcal{K}_{k+1}^{\delta^n}$$

and

$$(2.29) \quad \Delta : \text{span}_{\mathcal{K}_k^{\delta^n}} \{d\lambda | \lambda \in \mathcal{K}_k^{\delta^n}\} \rightarrow \text{span}_{\mathcal{K}_{k+1}^{\delta^n}} \{d\eta | \eta \in \mathcal{K}_k^{\delta^{n+1}}\} \subset \text{span}_{\mathcal{K}_{k+1}^{\delta^n}} \{d\gamma | \gamma \in \mathcal{K}_{k+1}^{\delta^n}\}$$

can be defined as in (2.17) and (2.18), respectively.

The final step of the analysis is to reduce the study of the chain $\mathcal{E}_0^+ \subset \mathcal{E}_1^+ \subset \dots$ to that of a related chain to which the proof technique of §2.2, that is, Lemmas 2.2–2.5,

can be applied with only minor modifications. As in §2.2, it is necessary to slightly modify the definition of \mathcal{E}_k^+ without changing its dimension:

$$(2.30) \quad \mathcal{E}_k^+ := \text{span}_{\mathcal{K}_{n+k}} \{dx, du[0], \dots, du[n-1], dy[n], \dots, dy[n+k]\},$$

for all $k \geq 0$, and

$$(2.31) \quad \mathcal{E}_{-1}^+ := \text{span}_{\mathcal{K}_{n-1}} \{dx, du[0], \dots, du[n-1]\}.$$

Note that $\{dx[n]\} \subset \mathcal{E}_{-1}^+$. Let $L \subset \{1, \dots, n\}$ be such that $\{dx_i[n] | i \in L\}$ is a basis for $\text{span}_{\mathcal{K}_{n-1}} \{dx[n]\}$, and let \mathcal{W} be such that

$$(2.32) \quad \mathcal{E}_{-1}^+ = \text{span}\{dx_i[n] | i \in L\} \oplus \mathcal{W}.$$

Introduce

$$(2.33) \quad \begin{aligned} \tilde{\mathcal{S}}_k &:= \text{span}_{\mathcal{K}_{n+k}} \{dx_i[n], dy[n], \dots, dy[n+k] | i \in L\}, \\ \tilde{\mathcal{S}}_{-1} &:= \text{span}_{\mathcal{K}_{n-1}} \{dx_i[n] | i \in L\}. \end{aligned}$$

Since $y[n+j](x, u[0], \dots, u[n+j]) = y[j](x[n], u[n], \dots, u[n+j])$ for $j \geq 0$, it follows that

$$(2.34) \quad \mathcal{E}_k^+ = \tilde{\mathcal{S}}_k \oplus \mathcal{W}, \quad k \geq 1.$$

Hence, for $k \geq 0$,

$$(2.35) \quad \sigma_k^+ = \dim \tilde{\mathcal{S}}_k - \dim \tilde{\mathcal{S}}_{k-1}.$$

The reason for doing all of this is that the generators of $\tilde{\mathcal{S}}_k$, for $k \geq 1$, are elements of $\{d\lambda | \lambda \in \mathcal{K}_k^{\delta^n}\}$. Hence, letting

$$(2.36) \quad \begin{aligned} \mathcal{S}_k &:= \text{span}_{\mathcal{K}_k^{\delta^n}} \{dx_i[n], dy[n], \dots, dy[n+k] | i \in L\}, \\ \mathcal{S}_{-1} &:= \text{span}_{\mathcal{K}_{n-1}} \{dx_i[n] | i \in L\}, \end{aligned}$$

it follows that, for $k \geq 0$,

$$(2.37) \quad \sigma_k^+ = \dim \mathcal{S}_k - \dim \mathcal{S}_{k-1}.$$

Moreover, even though $\mathcal{S}_k \subset \mathcal{S}_{k+1}$ is *not* a subspace of \mathcal{S}_{k+1} ,

$$(2.38) \quad \dim \text{span}_{\mathcal{K}_{k+1}^{\delta^n}} \{\mathcal{S}_k\} = \dim \mathcal{S}_k,$$

and therefore, the proofs of the obvious modifications of Lemmas 2.2–2.5 for the chain $\mathcal{S}_0 \subset \mathcal{S}_1 \subset \dots$ go through with only minor changes, which will not be repeated here.

THEOREM 2.8. *For all integers $k \geq n$, $\dim \mathcal{E}_k^+ - \dim \mathcal{E}_{k-1}^+ = \dim \mathcal{E}_n^+ - \dim \mathcal{E}_{n-1}^+$; that is, $\rho_k^+ - \rho_{k-1}^+ = \rho_n^+ - \rho_{n-1}^+$. Moreover, whenever (2.1) is generically submersive, the ordered lists $\{\sigma_0, \sigma_1, \dots, \sigma_n\}$ and $\{\sigma_0^+, \dots, \sigma_n^+\}$ are equal.*

The last part of the theorem follows from the fact that when (2.1) is generically submersive, the index set L in (2.32) is equal to $\{1, \dots, n\}$. Then, since $y[n+j](x, u[0], \dots, u[n+j])$ can be expressed as $y[j](x[n], u[n], \dots, u[n+j])$, \mathcal{E}_k and \mathcal{S}_k are naturally isomorphic under $x \rightarrow x[n], u[0] \rightarrow u[n], \dots, u[k] \rightarrow u[n+k]$.

3. Further characterizations of the rank and structure at infinity. This section and the rest of the paper will concentrate on generically submersive systems. Similar results, as per the development of §2.3, can be stated for the general case.

3.1. Jacobian matrices. The goal here is to provide a computationally convenient means of evaluating the rank ρ^* . The same result is also useful for showing the invariance of ρ^* under the action of invertible (static or dynamic) state variable feedback.

Following [14], which, in turn, was based upon [23], consider the Jacobian matrices

$$(3.1) \quad J_k(x, u[0], \dots, u[k]) := \frac{\partial(y[0], \dots, y[k])}{\partial(u[0], \dots, u[k])},$$

for $0 \leq k \leq n$, and their associated ranks

$$(3.2) \quad R_k := \text{rank}_{\mathcal{K}} J_k.$$

Note that the matrices J_k can be evaluated symbolically; their ranks can be evaluated numerically since the rank over \mathcal{K} is the same as the generic rank considered in [23].

Applying arguments identical to those used in [4, §2.1] results in the following relation between the integers ρ_k and R_k .

PROPOSITION 3.1. *For each $0 \leq k \leq n$, $\rho_k = n + R_k$. Hence, if (2.1) is generically submersive, then $\rho^* = R_n - R_{n-1}$.*

A quite different way of obtaining a result similar to the first part of Proposition 3.1 is given in [6].

Consider now a discrete-time linear system

$$(3.3) \quad \begin{aligned} x[k+1] &= Ax[k] + Bu[k], \\ y[k] &= Cx[k] + Du[k]. \end{aligned}$$

Then the Jacobian matrix J_k is given by the usual Toeplitz matrix

$$(3.4) \quad J_k = \begin{bmatrix} D & 0 & 0 \\ CB & D & \\ CAB & CB & \vdots \\ \vdots & & 0 \\ CA^{k-1}B & CA^{k-2}B & \dots & D \end{bmatrix}.$$

The results of [21] and [26] in conjunction with Proposition 3.1 justify the terminology adopted in §2.2 concerning the rank and structure at infinity of a nonlinear system.

The following is the analogue of [10, III.B.2. Proposition].

COROLLARY 3.2. *In the case of a linear system, the rank ρ^* defined by (2.10) agrees with the classical rank of the transfer matrix. Moreover, the list of integers $\{\sigma_0, \dots, \sigma_n\}$ defined in (2.15) is precisely the structure at infinity as it is normally defined on the basis of the transfer matrix [21], [26].*

REMARK 3.3. For a linear system, it is easy to verify that the lists $\{\sigma_k^+\}$ and $\{\sigma_k\}$ coincide, whether or not the system is generically submersive.

3.2. A related chain of subspaces. Related to the chain $\mathcal{E}_0 \subset \mathcal{E}_1 \subset \dots$ is the chain $\mathcal{H}_0 \subset \mathcal{H}_1 \subset \dots$ defined solely in terms of the output [4]:

$$(3.5) \quad \mathcal{H}_k = \text{span}_{\mathcal{K}_k} \{dy[0], \dots, dy[k]\}.$$

It also can be used to determine the rank of the system, and this will be important for making contact with the fundamental work of [10].

THEOREM 3.4. *Suppose that (2.1) is generically submersive. For all integers $k \geq n$, $\rho^* = \dim \mathcal{H}_k - \dim \mathcal{H}_{k-1}$.*

Proof. By Lemma 2.5, for $k \geq n$ $\rho^* \geq \dim \mathcal{H}_k - \dim \mathcal{H}_{k-1}$. On the other hand, for $1 \leq j$, $\mathcal{H}_j = \mathcal{H}_{j-1} + \text{span}\{dy[j]\}$, $\mathcal{E}_j = \mathcal{E}_{j-1} + \text{span}\{dy[j]\}$ and $H_0 \subset \mathcal{E}_0$. Thus, $\dim \mathcal{H}_j - \dim \mathcal{H}_{j-1} \geq \dim \mathcal{E}_j - \dim \mathcal{E}_{j-1}$. \square

3.3. Remarks on the inversion algorithm. The importance of the inversion algorithm of Singh [27], which is an extension to nonlinear continuous-time systems of the well-known algorithm of Silverman [26], need not be underlined here. The algorithm has also been used in the study of discrete-time nonlinear systems [18] and [19], but always expressed in a form involving the implicit function theorem. Consequently, the results of the algorithm can be difficult to interpret unless one remains in a neighborhood of an equilibrium point. This problem can be removed by working at the level of the differentials of the outputs, which linearizes the computations and allows the analysis of [4] to be carried through to the discrete-time setting. Since the algorithm in the form we will use it has already appeared in several publications for continuous-time systems [3], [4], [16], the basic idea will only be sketched here by giving the first steps of the algorithm. Establishing the validity and convergence properties of the algorithm is quite easy using the analysis of §2.2.

It is assumed that (2.1) is generically submersive; an extension to general systems can be envisioned along the lines of §2.3.

Step 0. Calculate $dy[0]$ and write it as

$$(3.6) \quad dy[0] = a_0(x, u[0])dx + b_0(x, u[0])du[0] .$$

Define

$$(3.7) \quad s_0 := \text{rank}_{\mathcal{K}_0} b_0 .$$

Permute, if necessary, the components of y so that the first s_0 rows of b_0 are linearly independent. Decompose y so that

$$(3.8) \quad dy[0] = \begin{bmatrix} d\tilde{y}_0[0] \\ d\hat{y}_0[0] \end{bmatrix} = \begin{bmatrix} \tilde{a}_0 \\ \hat{a}_0 \end{bmatrix} dx + \begin{bmatrix} \tilde{b}_0 \\ \hat{b}_0 \end{bmatrix} du[0],$$

where \tilde{y}_0 has s_0 rows. Since the rows of \hat{b}_0 are \mathcal{K}_0 -dependent on the rows of \tilde{b}_0 , there exists a matrix $M_0(x, u[0])$ with entries in \mathcal{K}_0 such that

$$(3.9) \quad \hat{b}_0 = M\tilde{b}_0,$$

and thus,

$$(3.10) \quad \begin{aligned} d\hat{y}_0[0] &= \hat{a}_0 dx + M_0\{d\tilde{y}_0[0] - \tilde{a}_0 dx\} \\ &=: \bar{a}_0 dx + \bar{b}_0 d\tilde{y}_0[0]. \end{aligned}$$

End of Step 0.

Step 1. Compute

$$(3.11) \quad \begin{aligned} d\hat{y}_0[1] &= (\delta\bar{a}_0) dx[1] + (\delta\bar{b}_0) d\tilde{y}_0[1] \\ &=: a_1(x, u[0], u[1]) dx + b_1(x, u[0], u[1]) du[0] + c_1(x, u[0], u[1]) d\tilde{y}_0[1]. \end{aligned}$$

Define

$$(3.12) \quad s_1 := \text{rank}_{\mathcal{K}_1} \begin{bmatrix} \tilde{b}_0 \\ b_1 \end{bmatrix}$$

and repeat the basic operations of Step 1; see [3], [4], and [16], for example.

The validation of the steps in the algorithm is achieved by noting that it produces a basis for $\mathcal{E}_0 \subset \mathcal{E}_1 \subset \dots$, and thus, by Lemma 2.2, since $\{dx, d\tilde{y}_0[0]\}$ is a linearly independent set, so is $\{dx, d\tilde{y}_0[0], d\tilde{y}_0[1]\}$, etc. Its convergence in no more than n steps follows from Theorem 2.1.

A similar connection with the interesting work of [24] on dynamic feedback solutions to the noninteracting control problem could be pursued also along the lines already clearly established in [4].

4. Invertibility. A linear system is usually said to be right-invertible if the rank of its transfer matrix is equal to the number of output components, and left-invertible if its rank equals the number of input components. Systemically, right-invertibility means that by a proper choice of the initial condition *and* input sequence, any output sequence can be generated; that is, the map from initial conditions and inputs is onto $Y^\infty = Y \times Y \times \dots$, the space of all output sequences. Left-invertibility is equivalent to injectivity of the map from inputs to outputs, for a fixed initial condition.

In the case of nonlinear systems, though such global notions of invertibility are attractive, simple examples show the difficulty of trying to say anything intelligent about them; hence, we are led to localizing the concepts. Following [3], for $k \geq 0$, let $H_k : X \times U^{k+1} \rightarrow Y^{k+1}$ be the map that sends $(x, u[0], \dots, u[k])$ to $(y[0], \dots, y[k])$, and let $E_k : X \times U^{k+1} \rightarrow X \times Y^{k+1}$ by $(x, u[0], \dots, u[k]) \mapsto (x, y[0], \dots, y[k])$.

DEFINITION 4.1. The system (2.1) is *almost everywhere locally surjective*, if, for every $k \geq 0$, the image of H_k has nonempty interior. The system is *almost everywhere locally injective*, if, for every $k \geq 0$, there exists an open and dense subset \mathcal{O}_k of $X \times U^{k+1}$ with the property that, for each point $p = (x, u[0], \dots, u[k]) \in \mathcal{O}_k$ there exists an open neighborhood of p , $\tilde{\mathcal{O}}_k(p)$, and an analytic insertion³ $i_k : \tilde{\mathcal{O}}_k(p) \rightarrow X \times U^{k+1} \times U^n$ such that if $p_1, p_2 \in \tilde{\mathcal{O}}_k(p)$ and $E_{n+k}(i_k(p_1)) = E_{n+k}(i_k(p_2))$, then $p_1 = p_2$.

These properties can be characterized as follows.

THEOREM 4.2. Assume that the nonlinear system (2.1) is generically submersive. Then the system is almost everywhere locally surjective if and only if any one of the following equivalent conditions is satisfied:

- (a) for all $k \geq 0$, $\dim \text{span}\{dy[0], \dots, dy[k]\} = (k+1)\mu$;
- (b) $\dim \text{span}\{dy[0], \dots, dy[n]\} = (n+1)\mu$;
- (c) $\rho^* = \mu$; i.e., the rank of the system equals the number of output components.

The system is almost everywhere locally injective if, and only if, any one of the following equivalent conditions is satisfied:

- (d) for all $k \geq 0$, $\{du[0], \dots, du[k]\} \subset \text{span}\{dx, dy[0], \dots, dy[k+n]\}$
- (e) $du[0] \subset \text{span}\{dx, dy[0], \dots, dy[n]\}$;
- (f) $\rho^* = m$; i.e., the rank of the system equals the number of input components.

Proof. (a) \implies (b) is immediate. (b) \implies (c) is given by Theorem 3.4. (c) \implies (a) follows from the same kind of reasoning employed in proving Lemmas 2.4 and 2.5 and is not repeated here. It suffices to show that (a) is equivalent to almost everywhere

³ That is, if τ_k represents that natural projection of $X \times U^{k+1} \times U^n$ onto $X \times U^{k+1}$, then $\tau_k \circ i_k|_{\tilde{\mathcal{O}}_k(p)}$ is the identity.

local surjectivity. The image of H_k has nonempty interior if, and only if, there exists an open set of points where the rank of H_k over the reals equals $(k+1)\mu$. This is equivalent to H_k having rank $(k+1)\mu$ over \mathcal{K}_k , which is equivalent to (a). Turning to almost everywhere local injectivity, (d) \implies (e) is evident. (e) \iff (f) is Corollary A.2 in Appendix A. It is now shown that (e) \implies (d): Applying Δ to both sides of (e) yields

$$du[1] \subset \text{span}\{dx[1], dy[1], \dots, dy[n+1]\}.$$

Hence, adding $\text{span}\{dx, du[0], dy[0]\}$ to the right-hand side,

$$\begin{aligned} du[1] &\subset \{dx, du[0], dy[0], dy[1], \dots, dy[n+1]\} \\ &\subset \text{span}\{dx, dy[0], \dots, dy[n+1]\}, \end{aligned}$$

where (e) has been used in the last step. The remainder of a proof by induction is clear. To finish up, it suffices now to show that (d) is equivalent to almost everywhere local injectivity. Without loss of generality, it can be assumed that the neighborhood $\tilde{\mathcal{O}}_k$ is such that $E_{n+k} \circ i_k : \tilde{\mathcal{O}}_k \rightarrow X \times Y^{n+k+1}$ has constant rank. Then almost everywhere local injectivity means that this rank equals $n + (k+1) \cdot m$. This is equivalent to $\text{rank}_{\mathcal{K}_{n+k}} (\partial(y[0], \dots, y[n+k]) / \partial(u[0], \dots, u[k])) = (k+1)m$, which is equivalent to (d). \square

Remark. A quite different approach to invertibility is taken in [10]; the results of the next section show that, in the case of polynomial systems, the two approaches coincide.

5. Difference algebra and the transformal transcendence degree. The purpose of this section is to prove that the rank ρ^* defined in (2.10), when specialized to systems whose right-hand side depends polynomially on x and u (more precisely, on their components), corresponds to the transformal transcendence degree used in [10].

Consider a system

$$(5.1) \quad \begin{aligned} \Sigma_{P,Q} \quad x[k+1] &= P(x[k], u[k]) \\ y[k] &= Q(x[k], u[k]), \end{aligned}$$

where $P : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $Q : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^\mu$ are such that each of their components is a polynomial of x and u , with coefficients in \mathbb{R} . This system is clearly analytic, so the analysis of §2 applies. For the purpose of clarity in presenting the results, it will be assumed that (5.1) is generically submersive. An extension to rational systems could also be undertaken.

The following definition, adapted from [10], should actually be *derived* by constructing the difference field associated to $\Sigma_{P,Q}$ and then applying the definition used in [10].

DEFINITION 5.1. The transformal transcendence degree of $\Sigma_{P,Q}$, denoted $d^0(\Sigma_{P,Q})$ equals the maximal number of components of y , say $\{y_{i_1}, \dots, y_{i_\rho}\}$, such that for any $k \geq 0$ there does not exist any nontrivial polynomial π with coefficients in \mathbb{R} such that

$$\pi(y_{i_1}[0], \dots, y_{i_\rho}[0], \dots, y_{i_1}[k], \dots, y_{i_\rho}[k]) = 0.$$

In other words, for any $k \geq 0$, $y_{i_1}[0], \dots, y_{i_\rho}[0], \dots, y_{i_1}[k], \dots, y_{i_\rho}[k]$ viewed as polynomials of $x, u[0], \dots, u[k]$ are algebraically independent.

For those readers familiar with Kähler differentials, the equality $\rho^* = d^0(\Sigma_{P,Q})$ is immediate from Theorem 3.4. For the benefit of other readers, an independent, straightforward proof is given. Lemma 5.2, which follows, is well-known, though hard to find in the form presented (cf. [28]).

Let $v = (v_1, \dots, v_r)$ be an r -tuple of indeterminants, let $\mathbb{R}[v]$ denote the ring of polynomials of (v_1, \dots, v_r) with coefficients in \mathbb{R} and let $\mathbb{R}(v)$ be the corresponding field of rational functions. Define a vector space over $\mathbb{R}(v)$ by $V := \text{span}\{dv_1, \dots, dv_r\}$ and define the mapping $d : \mathbb{R}(v) \rightarrow V$ by

$$(5.2) \quad d\left(\frac{p(v)}{q(v)}\right) := \frac{1}{(q(v))^2} \sum_{j=1}^r \left(\frac{\partial p(v)}{\partial v_j} q(v) - p(v) \frac{\partial q(v)}{\partial v_j} \right) dv_j$$

in the usual way; see [30, Chap. 5, §10.5] for how to define a differential calculus of rational functions without taking limits.

LEMMA 5.2. *A collection of polynomials $\{P_1, \dots, P_k\} \subset \mathbb{R}[v]$ is algebraically independent if, and only if, the set $\{dP_1, \dots, dP_k\}$ is linearly independent in $(V, \mathbb{R}(v))$.*

Proof. Suppose that $\{P_1, \dots, P_k\}$ is algebraically independent; then $k \leq r$. Assume first that $k = r$. $\{P_1, \dots, P_r\}$ is then a basis for $\mathbb{R}[v]$, and consequently, for each $1 \leq i \leq r$, there is a nontrivial polynomial $Q_i(\lambda_1, \dots, \lambda_{r+1})$ such that $Q_i(v_i, P_1, \dots, P_r) = 0$.

PROPOSITION 5.3. *For each $1 \leq i \leq r$, the polynomial $\partial Q_i / \partial \lambda_1$ is nontrivial.*

Proof. Suppose it is trivial. Then $Q_i(v_i, P_1, \dots, P_r) = Q_i(0, P_1, \dots, P_r) =: \overline{Q}_i(P_1, \dots, P_r)$. Thus, \overline{Q}_i must be a trivial polynomial because $\{P_1, \dots, P_r\}$ is algebraically independent. It follows that Q_i is a trivial polynomial, which contradicts its definition. \square

Continuing with the proof of Lemma 5.2, since $0 = Q_i(v_i, P_1, \dots, P_r)$,

$$(5.3) \quad 0 = d(Q_i(v_i, P_1, \dots, P_r)) = \frac{\partial Q_i}{\partial \lambda_1}(v_i, P_1, \dots, P_r) dv_i + \sum_{j=1}^r \frac{\partial Q_i}{\partial \lambda_{j+1}}(v_i, P_1, \dots, P_r) dP_j;$$

see [30, Chap. 5] for the chain rule. By Proposition 5.3, $\partial Q_i / \partial \lambda_1$ is nontrivial, and thus,

$$(5.4) \quad dv_i = \sum_{j=1}^r k_{ij} dP_j,$$

where

$$k_{ij} := \left(\frac{\partial Q_i}{\partial \lambda_1}(v_i, P_1, \dots, P_r) \right)^{-1} \left(\frac{\partial Q_i}{\partial \lambda_{j+1}}(v_i, P_1, \dots, P_r) \right) \in \mathbb{R}(v).$$

Thus, $\text{span}\{dv_1, \dots, dv_r\} \subset \text{span}\{dP_1, \dots, dP_r\}$, proving the linear independence of $\{dP_1, \dots, dP_r\}$.

If $k < r$, then there exist P_{k+1}, \dots, P_r such that $\{P_1, \dots, P_r\}$ is a basis for $\mathbb{R}(v)$. From the above, $\{dP_1, \dots, dP_r\}$ is linearly independent, and therefore, so must be $\{dP_1, \dots, dP_k\}$.

To prove the other direction of the lemma, suppose that $\{P_1, \dots, P_k\}$ is algebraically dependent. Then there exists a nontrivial polynomial $Q(\lambda_1, \dots, \lambda_k)$ with coefficients in \mathbb{R} such that $Q(P_1, \dots, P_k) = 0$. Hence, $0 = \sum_{j=1}^k \partial Q / \partial \lambda_j (P_1, \dots, P_k) dP_j$,

proving that $\{dP_1, \dots, dP_k\}$ is a linearly dependent set in $(V, \mathcal{R}(v))$. This completes the proof of Lemma 5.2. \square

The constructions and results of §2 hold clearly for polynomial systems (5.1) with the field \mathcal{K} replaced⁴ by $\mathcal{R}[x, u[0], \dots, u[n]]$. This observation, combined with Lemma 5.2 and Theorem 3.4, yields the following result.

THEOREM 5.4. *For the polynomial system (5.1), $\rho^* = d^0(\Sigma_{P,Q})$.*

Remark. Lemma 5.2 can be equivalently stated as: The following two conditions are equivalent:

(a) There exists a nontrivial polynomial π such that $\pi(P_1, \dots, P_r) = 0$

(b) The set $\{dP_1, \dots, dP_r\}$ is linearly dependent in $(V, \mathcal{R}(v))$

The implication (a) \implies (b) remains true with π, P_1, \dots, P_r replaced by analytic functions. However, the converse is then true only *locally*, and even then only on subsets where certain constant dimensional conditions are met; indeed, this is the well-known Rank Theorem.

6. Relation to generic observation fields. The chain of subspaces $\mathcal{E}_0 \subset \dots \subset \mathcal{E}_k \subset \dots$ is measuring how the input components are appearing in the outputs, assuming that the initial state is known. In a similar manner, one could study how the initial state components appear in the outputs, assuming that the inputs are known. This is called *observability* [15], and in the context of the formalism of this paper, could be studied via the chain $\mathcal{O}_0 \subset \dots \subset \mathcal{O}_k \subset \dots$, where

$$(6.1) \quad \mathcal{O}_k := \text{span}_{\mathcal{K}_k} \{du[0], \dots, du[k], dy[0], \dots, dy[k]\}.$$

Then, as in [4] and this paper, a connection could be established between the ranks of certain Jacobian matrices, the dimensions of the subspaces \mathcal{O}_k and/or the transcendence degree of a certain (differential) field. This (plus a whole lot more, such as the construction of a realization theory for polynomial input-output maps) was done by Sontag [28] in 1979 for a very general class of discrete-time polynomial systems (cf. his generic observation fields, Q_f^K). An extension to analytic systems would follow along the lines of [31]. More recent work on the analysis of the observability of continuous-time systems by algebraic means can be found in [5] and the references therein.

Appendix A.

Proof of Lemma 2.2. For the sake of the study of invertibility in §3, it is useful to prove a little more than is required by the lemma. The following notation is used only in Appendix A; it will help to keep the formulas concise. If M is a subset of \mathcal{E} , then $[M]$ denotes its *span* (see [20, p. 16]). If $\{v_1, \dots, v_s\}$ is a set of *linearly independent* elements in \mathcal{E} , this will be denoted by $\{v_1, \dots, v_s\}^*$. For example, let $I_0 \subset \{1, \dots, \mu\}$ be such that $\{dx, dy_{i_0}[0] | i_0 \in I_0\}$ is a basis for \mathcal{E}_0 . Then, this can be succinctly stated as $[\{dx, dy_{i_0}[0] | i_0 \in I_0\}^*] = \mathcal{E}_0$.

Suppose that for some $0 \leq k$, index sets $I_0 \subset I_1 \subset \dots \subset I_k \subset \{1, \dots, \mu\}$ and $\{1, \dots, m\} \supset J_0 \supset J_1 \supset \dots \supset J_k$ have been selected so that for each $0 \leq t \leq k$,

$$(A.1) \quad \mathcal{E}_t = [\{dx, dy_{i_r}[r] | 0 \leq r \leq t, i_r \in I_r\}^*]$$

$$(A.2) \quad [dx, du[0], \dots, du[t]] = [\{dx, dy_{i_r}[r], du_{j_{t-r}}[r] | 0 \leq r \leq t, i_r \in I_r, j_r \in J_r\}^*].$$

⁴ So that the proof of Lemma 2.4, which used the Rank Theorem, can be carried over, one must note that (2.21) holds for (5.1) over \mathcal{K} if, and only if, it holds over $\mathcal{R}[x, u[0], \dots, u[n]]$.

and

$$(A.3) \quad du[0] \subset \mathcal{E}_t \oplus [\{du_j[0] | j \in J_t\}].$$

That this is possible for $k = 0$ is obvious. It will first be shown that $I_{k+1} \supset I_k$ can be chosen so that (A.1) holds for $0 \leq t \leq k + 1$.

CLAIM A.1. *The set*

$$\left\{ dx, dy_{i_r}[r], dy_{i_k}[k+1], du_{j_k}[0], du_{j_{k-r}}[r+1] | 0 \leq r \leq k, i_r \in I_r, j_r \in J_r \right\}$$

is linearly independent.

Proof. To see how the arguments go, consider first (A.2) for $t = 0$ and apply Δ to both sides to obtain

$$[dx[1], du[1]] = [\{dx[1], dy_{i_0}[1], du_{j_0}[1] | i_0 \in I_0, j_0 \in J_0\}].$$

Adding $[dx, du[0]]$ to both sides yields

$$[dx, du[0], du[1]] = [\{dx, du[0], dy_{i_0}[1], du_{j_0}[1] | i_0 \in I_0, j_0 \in J_0\}].$$

Applying (A.2) for $t = 0$ results in

$$[dx, du[0], du[1]] = [\{dx, dy_{i_0}[0], du_{j_0}[0], dy_{i_0}[1], du_{j_0}[1] | i_0 \in I_0, j_0 \in J_0\}].$$

The independence of the vectors on the left-hand side implies the independence of those on the right-hand side by counting the number of elements. In particular, the vectors $\{dx, dy_{i_0}[0], dy_{i_0}[1], | i_0 \in I_0, \}$ are linearly independent, and thus, one can choose $I_1 \supset I_0$ such that (A.1) holds for $0 \leq t \leq 1$.

In general, consider (A.2) for $t = k$ and apply Δ to both sides to obtain

$$(A.4) \quad \begin{aligned} & [dx[1], du[1], \dots, du[k+1]] \\ &= [\{dx[1], dy_{i_r}[r+1], du_{j_{k-r}}[r+1] | 0 \leq r \leq k, i_r \in I_r, j_r \in J_r\}]. \end{aligned}$$

Add $[dx, du[0]]$ to both sides to obtain

$$\begin{aligned} & [dx, du[0], \dots, du[k+1]] \\ &= [\{dx, du[0], dy_{i_r}[r+1], du_{j_{k-r}}[r+1] | 0 \leq r \leq k, i_r \in I_r, j_r \in J_r\}^*]. \end{aligned}$$

The independence of the vectors on the left-hand side implies the independence of those on the right-hand side by counting the number of elements. Applying (A.2) first for $t = 0$, and then successively for $t = 1, \dots, t = k$ results in

$$(A.5) \quad \begin{aligned} & [dx, du[0], \dots, du[k+1]] \\ &= [\{dx, dy_{i_0}[0], du_{j_0}[0], dy_{i_r}[r+1], du_{j_{k-r}}[r+1] | 0 \leq r \leq k, i_r \in I_r, j_r \in J_r\}^*] \\ &= [\{dx, dy_{i_0}[0], dy_{i_1}[1], du_{j_1}[0], dy_{i_s}[s+1], du_{j_{k-r}}[r+1] | 0 \\ &\quad \leq r \leq k, 1 \leq s \leq k, i_s \in I_s, j_r \in J_r\}^*] \\ &= [\{dx, dy_{i_0}[0], dy_{i_1}[1], dy_{i_2}[2], du_{j_2}[0], dy_{i_s}[s+1], du_{j_{k-r}}[r+1] | 0 \\ &\quad \leq r \leq k, 2 \leq s \leq k, i_r \in I_r, j_r \in J_r\}^*] \\ &= \text{for each } 0 \leq t \leq k \\ &= [\{dx, dy_{i_0}[0], \dots, dy_{i_t}[t], du_{j_t}[0], dy_{i_s}[s+1], du_{j_{k-r}}[r+1] | 0 \\ &= \quad \leq r \leq k, t \leq s \leq k, i_r \in I_r, j_r \in J_r\}^*]. \quad \square \end{aligned}$$

Hence, we can choose $I_{k+1} \supset I_k$ so that (A.1) holds for each $0 \leq t \leq k+1$. It is next shown that $J_{k+1} \subset J_k$ can be chosen such that (A.2) and (A.3) hold. From (A.3) for $t = k$, since $\mathcal{E}_k \subset \mathcal{E}_{k+1}$, we deduce

$$(A.6) \quad du[0] \subset \mathcal{E}_{k+1} + \{du_j[0] | j \in J_k\}.$$

By the definition of I_{k+1} ,

$$(A.7) \quad \mathcal{E}_{k+1} = \mathcal{E}_k \oplus [\{dy_{i_k}[k+1] | i_k \in I_{k+1}\}^*] \oplus [\{dy_\ell[k+1] | \ell \in I_{k+1} \setminus I_k\}^*].$$

For each $\ell \in I_{k+1} \setminus I_k$, $dy_\ell[k] \in \mathcal{E}_k$, and thus,

$$(A.8) \quad \begin{aligned} dy_\ell[k+1] &\in [\{dx[1], dy_{i_0}[1], \dots, dy_{i_k}[k+1] | i_r \in I_r, 0 \leq r \leq k\}] \\ &\subset (\mathcal{E}_k \oplus [\{dy_{i_k}[k+1] | i_k \in I_k\}^*]) + [du[0]] \\ &\subset (\mathcal{E}_k \oplus [\{dy_{i_k}[k+1] | i_k \in I_k\}^*]) + [du[0] | j \in J_k], \end{aligned}$$

where the last inclusion is from (A.3) for $t = k$. Thus, $(|I_{k+1}| - |I_k|)$ elements of $\{du_j[0] | j \in J_k\}$ are not independent of \mathcal{E}_{k+1} . One can therefore choose $J_{k+1} \subset J_k$ such that

$$(A.9) \quad du[0] \subset \mathcal{E}_{k+1} \oplus [\{du_j[0] | j \in J_{k+1}\}^*]$$

and

$$(A.10) \quad |J_{k+1}| = |J_k| - (|I_{k+1}| - |I_k|).$$

To finish up, from (A.5) and (A.9) it follows that

$$(A.11) \quad \begin{aligned} [dx, du[0], \dots, du[k+1]] &\subset [\{dx, dy_{i_0}[0], \dots, dy_{i_{k+1}}[k+1], du_{j_{k+1}}[0], \dots, \\ &\quad du_{j_0}[k+1] | i_t \in I_t, j_{k+1-t} \in J_t, 0 \leq t \leq k+1\}]. \end{aligned}$$

Since the reverse inclusion is obviously true, one has equality. By counting the number of vectors on the right-hand side of (A.11), we obtain (A.2) for $t = k+1$. \square

From (A.1) and (A.2), respectively, it follows that

$$(A.12) \quad \rho_k = n + \sum_{i=1}^k |I_i|$$

and

$$(A.13) \quad n + (k+1)m = n + \sum_{i=1}^k |I_i| + \sum_{i=1}^k |J_i|,$$

yielding

$$(A.14) \quad \rho_k - \rho_{k-1} = m - |J_k|.$$

This combined with (A.3) proves the following result.

COROLLARY A.2. *Suppose that (2.1) is generically submersive. Then*

$$\dim \mathcal{E}_k - \dim \mathcal{E}_{k-1} = m \text{ if and only if } \{du_1[0], \dots, du_m[0]\} \subset \mathcal{E}_k .$$

Proof of Lemma 2.4. We can view $y[k]$, $0 \leq k \leq N$ as being an analytic function on $X \times U^{N+1}$. Let \mathcal{O} be an open subset of $X \times U^{N+1}$.

By the definition of I_0, \dots, I_n , for any $0 \leq N \leq n$,

$$(A.15) \quad dy_j[N] \in \text{span}_{\mathcal{K}_N} \{dx, dy_{i_0}[0], \dots, dy_{i_N}[N] | i_k \in I_k, \ 0 \leq k \leq N\} .$$

Due to analyticity, (A.15) is equivalent to

$$(A.16) \quad dy_j[N]|_{\mathcal{O}} \in \text{span}_{\mathbf{R}} \{dx, dy_{i_0}[0], \dots, dy_{i_N}[N] | i_k \in I_k, \ 0 \leq k \leq N\}|_{\mathcal{O}} ,$$

where the left-hand side is viewed as a one-form on \mathcal{O} , the right-hand side is viewed as an analytic codistribution on \mathcal{O} , and the span is taken pointwise; without loss of generality, it can be assumed that the codistribution has constant dimension. After possibly shrinking \mathcal{O} , the Rank theorem implies that on \mathcal{O} , $y_j[N]$ can be expressed as an analytic function of $(x, y_{i_0}[0], \dots, y_{i_N}[N] | i_k \in I_k, \ 0 \leq k \leq N)$. Repeating the above reasoning, (2.21) is equivalent to

$$(A.17) \quad dy_j[N]|_{\mathcal{O}} \in \text{span}_{\mathcal{K}_N} \{dy_j[0], \dots, dy_j[N-1], dy_{i_0}[0], \dots, dy_{i_N}[N] | i_k \in I_k, \ 0 \leq k \leq N\}|_{\mathcal{O}} .$$

Hence, from (A.16), (2.21) holds if

$$(A.18) \quad \frac{\partial y_j[N]}{\partial x} dx \in \text{span}_{\mathcal{K}_N} \left\{ \frac{\partial y_j[0]}{\partial x} dx, \dots, \frac{\partial y_j[N-1]}{\partial x} dx \right\} .$$

Because the right-hand side of (A.18) can be at most n -dimensional, there must exist an N , $1 \leq N \leq n$, such that this is the case. \square

Appendix B.

Proofs of Lemmas 2.6 and 2.7. The inclusion $M_{k+1} \subset M_k$, implies $d_{k+1} \leq d_k \ \forall k \geq 0$. Since $x[k] : \mathbf{R}^n \times (\mathbf{R}^m)^k \rightarrow \mathbf{R}^n$ is an analytic function, there exists an open and dense subset $V_k \subset \mathbf{R}^n \times (\mathbf{R}^m)^k$ on which $x[k]$ has constant \mathbf{R} -rank equal to d_k ; that is, for each point $p \in V_k$,

$$(B.1) \quad \text{rank}_{\mathbf{R}} \left[\frac{\partial x[k]}{\partial (x, u[0], \dots, u[k-1])} \right] (p) = d_k .$$

Let $N_k := x[k](V_k)$. Then, N_k is an immersed submanifold of \mathbf{R}^n , but since the resulting topology may be different than the subset topology of \mathbf{R}^n , this property is of little interest. More importantly, the implicit function theorem implies that, for each point $q \in N_k$, there exists an open subset $\mathcal{O}_k \subset V_k$ such that $x[k](\mathcal{O}_k)$ is a d_k -dimensional, embedded, analytic submanifold of \mathbf{R}^n and q is in the interior of $x[k](\mathcal{O}_k)$. Since V_k is dense in $\mathbf{R}^n \times (\mathbf{R}^m)^k$, N_k is dense in M_k in the subset topology.

Let $\bar{f}_k : x[k](\mathcal{O}_k) \times \mathbf{R}^m \rightarrow \mathbf{R}^n$ denote f restricted to $x[k](\mathcal{O}_k) \times \mathbf{R}^m$ (the subscript k is to note that \bar{f}_k depends on \mathcal{O}_k). Whenever $\mathcal{O}_k \subset V_k$ is such that $x[k](\mathcal{O}_k)$ is a d_k -dimensional embedded submanifold of \mathbf{R}^n , then \bar{f}_k is an analytic function. Consequently, it will have constant \mathbf{R} -rank on an open and dense subset of its domain

of definition, which is called its *generic rank* and is denoted as $\text{gen rank } \bar{f}_k$. From (2.23) and (2.24) it follows that

$$(B.2) \quad d_{k+1} = \text{gen rank } \bar{f}_k : x[k](\mathcal{O}_k) \times \mathbb{R}^m \rightarrow \mathbb{R}^n$$

since, if $p \in \mathcal{O}_k$ and $q := x[k](p)$, then $T_q x[k](\mathcal{O}_k)$, the tangent space of $x[k](\mathcal{O}_k)$ at the point q , satisfies

$$(B.3) \quad T_q x[k](\mathcal{O}_k) = \text{Image} \left[\frac{\partial x[k]}{\partial (x, u[0], \dots, u[k-1])} \right] (p).$$

With the above preliminaries completed, the proof of Lemma 2.6 can be given. If $d_n = 0$, the result is obvious; suppose, therefore, that $d_n > 0$. Then, there exists $0 \leq k \leq n-1$ such that $d_k = d_{k-1}$. As before, choose \mathcal{O}_k to be an open subset of V_k such that $x[k](\mathcal{O}_k)$ is an embedded d_k -dimensional submanifold of \mathbb{R}^n . Since $M_k \subset M_{k-1}$, and N_k and N_{k-1} are dense in M_k and M_{k-1} , respectively, the condition $d_k = d_{k-1}$ implies that $x[k](\mathcal{O}_k)$ is an embedded $(d_{k-1} = d_k)$ -dimensional submanifold of \mathbb{R}^n and $x[k](\mathcal{O}_k) \cap x[k-1](\mathcal{O}_{k-1})$ has nonempty interior. Therefore,

$$(B.4) \quad \begin{aligned} d_{k+1} &= \text{gen rank } \bar{f}_k : x[k](\mathcal{O}_k) \times \mathbb{R}^m \rightarrow \mathbb{R}^n \\ &= \text{gen rank } \bar{f}_{k-1} : x[k-1](\mathcal{O}_{k-1}) \times \mathbb{R}^m \rightarrow \mathbb{R}^n \\ &= d_k, \end{aligned}$$

where the fact that \bar{f}_k and \bar{f}_{k-1} are the restrictions of a common map f and have a common nonempty open set in their domain of definition entails the second equality. This completes the proof of Lemma 2.6.

Turning to Lemma 2.7, let $d := d_n$, which is then equal to d_k for all $k > n$ by Lemma 2.6. If $d = 0$, then Lemma 2.7 is immediate, so in the following it is supposed that $d > 0$. Let $\alpha = \alpha(x, u[0], \dots, u[k])$ be an element of \mathcal{R}_k . For any $r > 0$, $\delta^r(\alpha) = \alpha(x[r], u[r], \dots, u[k+r])$. Let $\tilde{\mathcal{O}}_r \subset V_r$ be an open set such that $x[r](\tilde{\mathcal{O}}_r)$ is a d_r -dimensional submanifold of \mathbb{R}^n . Then $\delta^r(\alpha) \neq 0$ if, and only if, α restricted to $(x[r](\tilde{\mathcal{O}}_r) \times (\mathbb{R}^m)^{k+1}) \neq 0$. When $r > n$, there exists $\tilde{\mathcal{O}}_n \subset V_n$ such that $x[n](\tilde{\mathcal{O}}_n) \times (\mathbb{R}^m)^{k+1} \cap x[r](\tilde{\mathcal{O}}_r) \times (\mathbb{R}^m)^{k+1}$ has nonempty interior. Thus, α restricted to $(x[r](\tilde{\mathcal{O}}_r) \times (\mathbb{R}^m)^{k+1}) \neq 0$ if, and only if, α restricted to $(x[n](\tilde{\mathcal{O}}_n) \times (\mathbb{R}^m)^{k+1}) \neq 0$. \square

Acknowledgment. The author has benefited greatly from conversations with M. Fliess and C.H. Moog on the subject of this paper, as well as from the insightful comments provided by the reviewers. This work was completed while the author was a visiting researcher at the Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France (Poste Rouge, Directeur de Recherche Associé). Professors P. Bertrand, M. Fliess, and F. Lamnabhi-Lagarigue are sincerely thanked for contributing in diverse ways to a very enjoyable and worthwhile *séjour*.

REFERENCES

- [1] LI CAO AND YU-FAN ZHENG, *On minimal compensators for decoupling control*, preprint, 1991.
- [2] M. D. DI BENEDETTO AND J. W. GRIZZLE, *An analysis of regularity conditions in nonlinear synthesis problems*, Lecture Notes in Control and Inform. Sci., Vol. 144, Springer-Verlag, Berlin, 1990, pp. 843–850.
- [3] ———, *Intrinsic notions of regularity for local inversion, output nulling and dynamic extension of non-square systems*, Control Theory Adv. Tech., 6 (1990), pp. 357–381.

- [4] M. D. DI BENEDETTO, J. W. GRIZZLE, AND C. H. MOOG, *Rank invariants of nonlinear systems*, SIAM J. Control Optim., 72 (1989), pp. 658–672.
- [5] S. DIOP AND M. FLIESS, *Nonlinear observability, identifiability and persistent trajectories*, in Proc. 30th IEEE Conference on Decision and Control, Brighton, England, December 1991, pp. 714–718.
- [6] S. EL ASMÍ AND M. FLIESS, *Formules d'inversion*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J.P. Gauthier, and I. Kupka, eds., Conf. Proc., Lyon, France, July 1990, Birkhäuser, Boston, MA.
- [7] ———, *Inversion formula for discrete-time systems*, IFAC Symposium on Nonlinear Control Systems Design, NOLCOS-92, Bordeaux, France, June 1992.
- [8] M. FLIESS, *A new approach to the noninteracting control problem in nonlinear systems theory*, in Proc. 23rd Allerton Conference, University of Illinois, Monticello, IL, 1985, pp. 123–129.
- [9] ———, *Generalized controller canonical forms for linear and nonlinear dynamics*, IEEE Trans. Automat. Control, AC-35, 9 (1990), pp. 994–1001.
- [10] ———, *Automatique en temps discret et algèbre aux différences*, Forum Mathematicum, 2 (1990), pp. 213–232.
- [11] ———, *Reversible linear and nonlinear discrete time dynamics* IEEE Trans. Automat. Control, AC-37, (1992), pp. 1144–1153.
- [12] M. FLIESS AND D. NORMAND-CYROT, *A group theoretic approach to discrete-time nonlinear controllability*, Proc. 20th IEEE Conf. Decision and Control, San Diego, 1981, pp. 551–557.
- [13] A. GLUMINEAU AND C. H. MOOG, *Essential orders and the nonlinear decoupling problem*, Internat. J. Control, 50 (1989), pp. 1825–1834.
- [14] J. W. GRIZZLE AND H. NIJMEIJER, *Zeros at infinity for nonlinear discrete-time systems*, Math. Syst. Theory, 19 (1986), pp. 79–93.
- [15] R. HERMAN AND H. K. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, 22 (1977), pp. 728–740.
- [16] A. J. C. HUIJBERTS, H. NIJMEIJER, AND L. L. M. VAN DER WEGEN, *Dynamic disturbance decoupling for nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 336–349.
- [17] B. JAKUBCZYK AND E. D. SONTAG, *Controllability of nonlinear discrete-time system: a Lie algebraic approach*, SIAM J. Control Optim., 28 (1990), pp. 1–33.
- [18] Ü. KOTTA, *Right-inverse of a discrete-time non-linear system*, Internat. J. Control, 51 (1990), pp. 1–9.
- [19] ———, *Local (finite-time) model matching of nonlinear discrete-time systems*, preprint, 1990.
- [20] D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [21] J. L. MASSEY AND M. K. SAIN, *Invertibility of linear time-invariant dynamic systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 141–149.
- [22] C. H. MOOG, *Nonlinear decoupling and structure at infinity*, Math. Control Signals Systems, 1 (1988), pp. 257–268.
- [23] H. NIJMEIJER, *Right-invertibility for a class of nonlinear control systems: a geometric approach*, Systems Control Lett., 7 (1986), pp. 125–132.
- [24] ———, *On dynamic decoupling and dynamic path controllability in economic systems*, J. Econom. Dynamics Control, 13 (1989), pp. 21–39.
- [25] D. NORMAND-CYROT, *Théorie and pratique des systèmes non linéaires en temps discret*, Thèse d'Etat, Université Paris-Sud, Orsay, 1983.
- [26] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 270–276.
- [27] S. N. SINGH, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 595–598.
- [28] E. D. SONTAG, *Polynomial Response Maps*, Lecture Notes in Control and Inform. Sci., 13, Springer, Berlin, 1979.
- [29] E. D. SONTAG AND Y. ROCHALEAU, *On discrete-time polynomial systems*, Nonlinear Anal.: Theory Meth. and Appl., 1 (1976), pp. 55–64.
- [30] B. L. VAN DER WAERDEN, *Modern Algebra*, Frederick Ungar Pub., New York, Vols. I and II, 1953.
- [31] Y. WANG AND E. D. SONTAG, *Realizations and I/O relations: The analytic case*, Proc. IEEE Conference on Decision and Control, Tampa, December 1989, pp. 1275–1280.
- [32] X.-H. XIA, *The essential structure of nonlinear systems*, preprint, 1989.

STOCHASTIC APPROXIMATION WITH AVERAGING OF THE ITERATES: OPTIMAL ASYMPTOTIC RATE OF CONVERGENCE FOR GENERAL PROCESSES*

HAROLD J. KUSHNER^{†‡} AND JICHUAN YANG^{†§}

Abstract. Consider the stochastic approximation algorithm

$$X_{n+1} = X_n + a_n g(X_n, \xi_n).$$

In an important paper, Polyak and Juditsky [*SIAM J. Control Optim.*, 30 (1992), pp. 838–855] showed that (loosely speaking) if the coefficients a_n go to zero slower than $O(1/n)$, then the averaged sequence $\sum_{i=1}^n X_i/n$ converged to its limit, at an optimum rate, for any coefficient sequence. The conditions were rather special, and direct constructions were used. Here a rather simple proof is given that results of this type are generic to stochastic approximation, and essentially hold any time that the classical asymptotic normality of the normalized and centered iterates holds. Considerable intuitive insight is provided into the procedure. Simulations have well borne out the importance of the method.

Key words. stochastic approximation, stochastic approximation with averaged iterates, rate of convergence for stochastic algorithms

AMS subject classifications. 62L20, 60F05, 62E25

1. Introduction. Consider the stochastic approximation (SA)

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

where $0 < a_n \rightarrow 0$, $\sum_n a_n = \infty$ and $\{\xi_n\}$ is some “noise” sequence. Suppose that for some θ , $X_n \rightarrow \theta$ either with probability one or weakly. Then, under appropriate conditions $(X_n - \theta)/\sqrt{a_n}$ converges in distribution to a normal random variable with mean zero and some covariance matrix V_0 . The matrix V_0 is often considered to be a measure of the “rate of convergence,” taken together with the scale factors or gains $\{a_n\}$.

Suppose that $a_n \rightarrow 0$ “slower” than $1/n$. In particular, suppose that

$$(1.2) \quad a_n/a_{n+1} = 1 + o(a_n).$$

Define

$$(1.3) \quad \bar{X}_n = \frac{1}{n} \sum_1^n X_i.$$

Then, in a sequence of fundamental papers, Polyak [1] and Polyak and Juditsky [2] showed that $\sqrt{n}(\bar{X}_n - \theta)$ converged in distribution to a normal random variable with mean zero and covariance V , where V was the smallest possible in an appropriate sense. V did not depend on $\{a_n\}$, provided that (1.2) held. This relaxation of conditions on $\{a_n\}$ allows it to be “relatively large” and is quite important in applications.

* Received by the editors July 1, 1991; accepted for publication (in revised form) February 28, 1992.

[†] Division of Applied Mathematics, Brown University, Box F, Providence, Rhode Island 02912.

[‡] The work of this author was supported by Air Force Office of Scientific Research grant AFOSR 89-0015, and Army Research Office grant ARO-DAAL 03-86K-0171.

[§] The work of this author was partially supported by National Science Foundation grant ECS-8913351.

Simulations have supported the theoretical conclusions and have shown the clear superiority of the use of averages of the type (1.3) over X_n directly. This superiority would not be true if a_n decreased as $O(1/n)$, and in this latter case, it was known for a long time that the asymptotic rates are the same for $\{\bar{X}_n\}$ and $\{X_n\}$. In the past, a great deal of attention was given to the problem of choosing optimal sequences $\{a_n\}$ via “adaptive” and generally unreliable means. The importance of this problem is now much reduced. The results in [1], [2] are similar in spirit to the approach of Ruppert [3] for a one-dimensional case.

The proofs in [1], [2] were by direct construction. They involved detailed expansions and estimates and made no use of prior results in SA. The function $g(\cdot)$ needed to be “smooth” and the conditions on the noise were restrictive, being essentially that $\{\xi_n\}$ were either i.i.d. or martingale differences, depending on the case. The conditions on the noise were weakened by Yin [4], [5] who allowed certain “mixing sequences,” but the proofs were still complicated and made no use of prior results in SA.

In this paper, it will be seen that a very simple use of prior results in SA allows us to get results of the above type under condition of considerable generality. In addition, the approach sheds more light on the reasons why averaged estimators such as (1.3) work well. In general, we use sums of the type (1.3), but where the lower index of summation goes to infinity as $n \rightarrow \infty$. We will work with several such averages.

In §2, we show that a very useful averaging result can be readily obtained via a weak convergence method under quite broad conditions. This result uses a “minimal window” of iterates, fewer than in (1.3), yet is quite useful in practice. Insight into the reasons why the averaging method works is obtained via an examination of a “two time scale SA” representation of (1.1), (1.3) in a simple case. The “window of averaging” is extended in §3.

2. The basic convergence theorem for the averaged iterates. Define the “interpolated time” $t_n = \sum_0^{n-1} a_i$, with $t_0 = 0$, and its “inverse” $m(t) = \max\{n : t_n \leq t\}$. For notational simplicity and without loss of generality, set $\theta = 0$. For each $n \geq 0$, define the interpolated processes $X^n(\cdot)$ and $U^n(\cdot)$ by

$$\left. \begin{aligned} X^n(t) &= X_{n+i} \\ U^n(t) &= X_{n+i}/\sqrt{a_{n+i}} \end{aligned} \right\} \text{ for } t \in [t_{n+i} - t_n, t_{n+i+1} - t_n], i \geq 0.$$

Let \Rightarrow denote weak convergence in the Skorohod topology on $D^r[0, \infty)$ [6], [7]. In Theorem 2.1, we will use the following assumption.

Assumption A2.1. There is a matrix G whose eigenvalues lie in the open left half plane and a positive definite symmetric matrix R_0 such that $X^n(\cdot) \Rightarrow$ zero process and $U^n(\cdot) \Rightarrow U(\cdot)$, where $U(\cdot)$ is the stationary solution to

$$(2.1) \quad dU = GUdt + R_0^{1/2}dw.$$

Comment on (A2.1). We prefer to state the condition in the form of (A2.1) since so many different sets of conditions imply (A2.1). Also, the main aim here is to show that a standard weak convergence result can be used to get an optimal rate of convergence under any of the sets of conditions which guarantee the usual limit result (A2.1). The references [8], [9], [11]–[13] contain various sets of conditions which guarantee (A2.1).

For $t > 0$, define $Z^n(\cdot)$ by

$$(2.2) \quad Z^n(t) = \frac{1}{\sqrt{t/a_n}} \sum_{i=n}^{n+t/a_n} X_i.$$

In sums of type \sum_{α}^{β} for real α, β , we always use the integer parts of α, β .

A basic convergence theorem.

THEOREM 2.1. *Assume (1.2) and (A2.1) and define $V = G^{-1}R_0(G')^{-1}$. For each t , $Z^n(t)$ converges in distribution to a random variable with mean zero and covariance $V_t = V + O(1/t)$.*

Proof. Define the processes

$$\tilde{Z}^n(t) = \frac{1}{\sqrt{t}} \int_0^t U^n(s) ds, \quad \tilde{Z}(t) = \frac{1}{\sqrt{t}} \int_0^t U(s) ds.$$

By the weak convergence in (A2.1), $\tilde{Z}^n(\cdot) \Rightarrow \tilde{Z}(\cdot)$. Define the covariance matrix $R(s) = EU(t)U'(t+s)$, where $U(\cdot)$ is the stationary solution to (2.1). Since $R(s) \rightarrow 0$ exponentially as $s \rightarrow \infty$,

$$\begin{aligned} \text{cov } \tilde{Z}(t) &= \frac{1}{t} \int_0^t \int_0^t R(s-\tau) ds d\tau \\ &= \int_{-\infty}^{\infty} R(s) ds + O(1/t), \end{aligned}$$

but $\int_{-\infty}^{\infty} R(s) ds = G^{-1}R_0(G^{-1})'$.

The basic result on the character of the averaged iterates is obtained by relating $Z^n(t)$ to $\tilde{Z}^n(t)$.

Write

$$(2.3) \quad \frac{a_n}{a_{n+i}} = 1 + \delta_{n,i}.$$

Then (1.2) implies that for any $t < \infty$,

$$(2.4) \quad \max\{i - n : 0 \leq t_i - t_n \leq t\} \cdot a_n/t \xrightarrow{n} 1.$$

Equation (2.4) will be heavily used. Note that (2.4) would not hold if $a_n = O(1/n)$.

Equation (2.4) follows from (1.2) as follows: Let $i > n$. Then $a_i/a_n = \prod_n^i (1 + o(a_j))$.

If $\sum_n^i a_j \leq t$, then the ratio goes to unity, uniformly in such i , as $n \rightarrow \infty$.

Using the "piecewise constant" definition of $U^n(\cdot)$, we have (modulo "end terms") for $i \geq n$

$$\begin{aligned} \sqrt{t}\tilde{Z}^n(t) &= \sum_{i:t_i-t_n \leq t} (X_i a_i^{-1/2}) a_i \\ &= \sum_{i:t_i-t_n \leq t} X_i (a_i^{1/2} - a_n^{1/2}) + a_n^{1/2} \sum_{i:t_i-t_n \leq t} X_i. \end{aligned}$$

(Alternatively, the sums can be written as $\sum_{m(t_n)}^{m(t_n+t)}$.) By the weak convergence of $U^n(\cdot)$ in (A2.1) and (2.4), the first sum on the right goes to zero in probability as

$n \rightarrow \infty$. By the same weak convergence and (2.4), the second sum on the right is asymptotically equivalent (in distribution) to

$$(2.5) \quad a_n^{1/2} \sum_{i=n}^{n+t/a_n} X_i.$$

This and the weak convergence $\tilde{Z}^n(t) \Rightarrow \tilde{Z}(t)$ yield the theorem. \square

Discussion of the theorem.

(a) On the optimality of the “rate of convergence.” Suppose that $U_n = X_n/\sqrt{a_n}$ converged in distribution to a normally distributed random variable \hat{U} with mean zero. It is common to consider the covariance of $\sqrt{a_n}\hat{U}$ as a “measure of the rate of convergence” or “asymptotic errors.” Then the best value of a_n is $O(1/n)$. Suppose that $a_n = A/n$, for A a positive definite matrix. Then, under appropriate conditions (see, e.g., [8], [9]), $U^n(\cdot) \Rightarrow \tilde{U}(\cdot)$, where $\tilde{U}(\cdot)$ is the stationary solution to

$$(2.6) \quad d\tilde{U} = \left(\frac{I}{2} + AG \right) \tilde{U} dt + AR_0^{1/2} dw,$$

where G and R_0 are as in (A2.1), and it is supposed that $(I/2 + AG)$ is a stable matrix. If we optimize the trace of the covariance matrix of (2.6) over A , we get the best value of A as

$$A = -G^{-1}.$$

With this value of A , the covariance of $\tilde{U}(0)$ is just the V used in Theorem 2.1. In this sense, the result in [1], [2] and of Theorem 2.1 is optimal. Let us note that the fact that $U^n(\cdot) \Rightarrow U(\cdot)$ satisfying (2.1), rather than (2.6), is of crucial importance. The integral of the correlation function of (2.6) depends on A . Again we emphasize that Theorem 2.1 requires that $a_n \rightarrow 0$ slower than $O(1/n)$. The result of Theorem 2.1 holds for some very complicated SAs, e.g., ones which arise due to distributed and asynchronous processing [13].

(b) The “window” of averaging. The value of t can be made as large as desired in (2.2), and can go to infinity slowly with n . More will be said about this in the next section. A two-sided average

$$\sqrt{\frac{a_n}{t_1 + t_2}} \sum_{n-t_1/a_n}^{n+t_2/a_n} X_i$$

can be used in lieu of (2.2) with the same results. For this case, the proof is nearly identical to the one given.

In (2.2), the “window” of the averaging is $O(1/a_n)$ as opposed to $O(n)$ in (1.3). Theorem 2.1 implies that the order $O(1/a_n)$ is the *smallest* which can be used. Suppose that $a_n = 1/n^\gamma$, $\gamma \in (0, 1)$. Then as $\gamma \rightarrow 0$, the minimal window of averaging decreases. Loosely speaking, for smaller rates of decrease of $\{a_n\}$, there is more “oscillation” of the iterates $\{X_n\}$ about the limit point, and less averaging is needed. This point will be supported by the following discussion of the singularly perturbed SA.

The theorem shows that the improvement, due to averaging, is a *natural property* of SA and is essentially a consequence of weak convergence of the normalized process $U^n(\cdot)$.

(c) Relationships between the use of (2.2) and a two time scale SA. For additional motivation and insight concerning the averaging method, let us rewrite the iteration for (X_n, U_n) as a two time scale or “singularly perturbed” SA. The following discussion is purely heuristic. Hence, for simplicity of presentation, we use a linear and one-dimensional model. For $A > 0$ and $\gamma \in (0, 1)$, define $\{X_n\}$ by

$$X_{n+1} = \left(1 - \frac{AG}{n^\gamma}\right)X_n + \frac{A\xi_n}{n^\gamma}.$$

Define $\bar{U}_n = \sum_1^n X_i/\sqrt{n}$. Then, putting the iterations for \bar{U}_n and X_n on the same time scale, we can write

$$(2.7a) \quad \frac{1}{n^{1-\gamma}}(X_{n+1} - X_n) = -\frac{AGX_n}{n} + \frac{A\xi_n}{n},$$

$$(2.7b) \quad \bar{U}_{n+1} - \bar{U}_n = -\frac{\bar{U}_n}{2n} \left(1 + O\left(\frac{1}{n}\right)\right) + \frac{X_{n+1}}{\sqrt{n+1}}.$$

\bar{U}_n is just \sqrt{n} times the averaged value $\sum_1^n X_i/n$, and a quantity whose asymptotic variance is of interest. Equation (2.7) can be viewed as a two time scale SA.

We can make a similar heuristic argument if \bar{U}_n is replaced by $\sqrt{t}Z^n(t)$. Define a new interpolation time $\tilde{t}_n = \sum_{i=1}^n 1/i$. Define the new continuous time interpolation $\tilde{X}^n(\cdot)$ by $\tilde{X}^n(t) = X_{n+i}$ on $[\tilde{t}_{n+i} - \tilde{t}_n, \tilde{t}_{n+i+1} - \tilde{t}_n)$, $i \geq 0$. In this new time scale, $\{X_n\}$ is “squeezed” or compressed more than it was in the $\{t_n\}$ scale and $\tilde{X}^n(\cdot)$ has a smaller correlation than $X^n(\cdot)$. This “smaller correlation” suggests that an averaging method will yield an improved result. Note that this two time scale effect doesn’t hold if $a_n = O(1/n)$.

A two time scale continuous parameter system that is loosely analogous to (2.7) is

$$(2.8) \quad \varepsilon dz^\varepsilon = A_{11}z^\varepsilon dt + dw_1$$

$$dx^\varepsilon = A_{22}x^\varepsilon dt + A_{12}z^\varepsilon dt + dw_2,$$

where ε is small. Under suitable stability conditions, [10, Chap. 10], $\int_0^t z^\varepsilon(s)ds$ converges weakly to a Wiener process. Hence, we might expect that, with (2.7), the function of t defined by

$$\frac{1}{\sqrt{n}} \sum_{i=0}^{nt} X_i$$

might also converge weakly to a Wiener process with covariance matrix V . Indeed, a similar result is proved in the next section.

Constant gain coefficients. Replace (1.1) by

$$(2.9) \quad X_{n+1}^\varepsilon = X_n^\varepsilon + \varepsilon g(X_n^\varepsilon, \xi_n), \quad \varepsilon > 0.$$

Define $X^\varepsilon(\cdot)$ and $U^\varepsilon(\cdot)$ by $X^\varepsilon(t) = X_n^\varepsilon, U^\varepsilon(t) = X_n^\varepsilon/\sqrt{\varepsilon}$ on $[n\varepsilon, n\varepsilon + \varepsilon)$. Suppose that there are $t_\varepsilon \xrightarrow{\varepsilon} \infty$ such that $U^\varepsilon(t_\varepsilon + \cdot) \Rightarrow U(\cdot)$, a stationary process which satisfies (2.1). Such results are proved in [11]. Fix $t > 0$ and let $t_i \geq 0$ such that $t = t_1 + t_2$. Define

$$Z^\varepsilon(t) = \sqrt{\frac{\varepsilon}{t}} \sum_{(t_\varepsilon - t_2)/\varepsilon}^{(t_\varepsilon + t_1)/\varepsilon} X_i.$$

Then, following the argument of Theorem 2.1 yields that for each t , $Z^\varepsilon(t)$ converges in distribution to a normally distributed random variable with mean zero and covariance $V + O(1/t)$.

A note on computation. In applications, averages of the type

$$\frac{1}{m_2(n) - m_1(n)} \sum_{m_1(n)}^{m_2(n)-1} X_i$$

are usually preferred to (1.3), where $m_i \rightarrow \infty$ and $m_2(n) - m_1(n) \rightarrow \infty$ as $n \rightarrow \infty$. Such averages cannot, in general, be computed recursively. However, memory requirements can be reduced by appropriate grouping of the iterates and selection of the times of updating.

3. Increasing the window of averaging in (2.2). In this section, we will see how fast we can let $t \rightarrow \infty$ and prove the above assertion concerning convergence to Wiener process.

For a sequence $n \geq q_n \rightarrow \infty$, define $M^n(t)$ by

$$(3.1) \quad M^n(t) = \frac{1}{\sqrt{q_n}} \sum_{i=n}^{n+q_n t} X_i.$$

We could use $\sum_{i=n-q'_n t}^{n+q''_n t}$ in (3.1) where $q''_n + q'_n = q_n$ and $q_n > \epsilon n$ for some $\epsilon > 0$, with the same end results. This is, in fact, a practical case since we would often wish to delete some initial fraction of the iterates from the averaging. To relate this last form to (1.3), let $t = 1$, $q_n = n$, $q'_n = (1 - \alpha)n$, $\alpha < 1$. Then we get

$$\frac{1}{n} \sum_{i=\alpha n}^{(1+\alpha)n} X_i.$$

Theorem 2.1 dealt with the case $q_n = O(1/a_n)$. In this section we need to assume that

$$(A3.1) \quad q_n a_n^{3/2} \rightarrow 0, \quad q_n a_n \rightarrow \infty, \quad q_n \leq k_0 n \quad \text{for some } k_0 < \infty.$$

Thus if $q_n = n$ and $a_n = 1/n^\gamma$, then $\gamma \in (2/3, 1)$ is needed. The result in [2], [3] required only $\gamma \in (1/2, 1)$, but our conditions on the noise and dynamics are more general. Under stronger conditions, the method of the following theorem can be carried through with $\gamma \in (1/2, 1)$. In the next section, it will be shown that $M^n(\cdot) \Rightarrow w(\cdot)$, a Wiener process with covariance matrix Vt , thus supporting the assertion at the end of the last section. In order to extend the "window of averaging" beyond t/a_n , we need $q_n a_n \rightarrow \infty$.

Discussion of the noise processes. Two types of noise processes are considered. For the first, the sequence $\{\xi_n\}$ is “exogenous.” Loosely speaking, the evolution of $\{X_n\}$ does not affect $\{\xi_n\}$. For the second, or “state dependent” noise, (X_n, ξ_n) is jointly Markov. There is a transition function $p(\xi, \cdot | x)$ such that $P\{\xi_{n+1} \in A | \xi_n = \xi, X_n = x\} = p(\xi, A | x)$. For each n and x , define the Markov process $\{\xi_j(x), j \geq n\}$ with initial condition $\xi_n(x) = \xi_n$ and transition function $p(\xi, \cdot | x)$. Such models were introduced in [11], [14], and also used in [9], [12].

The following additional conditions will be used. Let E_n denote the expectation conditioned on $\{X_i, i \leq n, \xi_i, i < n\}$.

Condition A3.2. There is a continuously differentiable “centering” function $\bar{g}(\cdot)$ such that with the definition $\psi_j(x) = g(x, \xi_j) - \bar{g}(x)$ (for the exogenous noise) and $\psi_j(x) = g(x, \xi_j(x)) - \bar{g}(x), j \geq n$ (for the state dependent noise) we have for each n and x ,

$$(3.2) \quad \sum_{j=n}^{\infty} a_j E_n \psi_j(x) = O(a_n)$$

where $O(a_n)$ is uniform in n, ω, x (where ω is the canonical point of the sample space), and $\psi_j(x)$ is bounded. In (3.2), for the state dependent noise case, the initial condition of $\{\xi_j(x), j \geq n\}$ is $\xi_n(x) = \xi_n$, following the usage in [12], [14].

Condition A3.3. $\bar{g}(x) = Gx + \delta g(x)$, where G has its eigenvalues in the open left half plane and $|\delta g(x)| = O(|x|^2)$.

Condition A3.4. $\sum_{j=n}^{\infty} a_j E_n [\psi_j(x) - \psi_j(y)] = O(a_n) |y - x|, |g(x, \xi)| \leq O(1) [|x| + 1]$.

Comments on the conditions (A3.2), (A3.4). Such conditions were initially introduced in [11], [15] and have been used in many of the other references; e.g., [4], [5], [9], [11], [12], [13]. They are essentially conditions on the “mixing rate” of the processes. A simple example of (3.2) is where the noise is “exogenous,” $\bar{g}(x) = Eg(x, \xi)$ is smooth, $\{\psi_j(x)\}$ is bounded and $\{\xi_n\}$ satisfies a mixing condition with a sufficiently fast mixing rate. Many specific examples are shown in [9], [12]. In [9], the sum in (A3.2) is the solution to the Poisson equation, and then (A3.4) is a condition on the smoothness of that solution. For the state dependent noise case, the transition kernel $p(\xi, \cdot | x)$ often assures that $\int g(x, \xi') p(\xi, d\xi' | x)$ is smooth enough so that (A3.4) holds [11].

We could replace $O(a_n)$ in (A3.2) by $O(a_n) \hat{g}(x)$ where $\hat{g}(x)$ has an appropriate growth rate, but we prefer to keep the development simple.

Stability of (1.1). A main problem in extending the window of averaging in Theorem 2.1 concerns the tightness of $\{X_n / \sqrt{a_n}\}$. Such results are basic to the proofs of (A2.1). In fact, given such tightness, straightforward averaging methods can often be used to get (A2.1). Theorem 3.1 requires the following bounds in Lemma 3.1, and to get that the following stability condition will be used. Recall that we use the assumption that the limit point is $\theta = 0$ for notational convenience and without loss of generality.

Condition A3.5. There is a nonnegative continuous function $V(\cdot)$ whose first and second mixed partial derivatives exist and are continuous. For some positive definite symmetric matrix A and some $\gamma > 0, K < \infty$,

$$V(x) = x' A x + o(|x|^2),$$

$$V'_x(x) \bar{g}(x) \leq -\gamma V(x), \quad |V_x(x)|^2 \leq K V(x),$$

$$|g(x, \xi)|^2 \leq K(V(x) + 1).$$

and $V_{xx}(x)$ is uniformly bounded.

Let \mathcal{F}_m denote the minimal σ -algebra which measures $\{X_i, i \leq m; \xi_i, i < m\}$.

LEMMA 3.1. Assume (A3.2)–(A3.5). Then $\{E|X_n|^4/a_n^2\}$ is bounded. Let $k < \infty$. For any \mathcal{F}_m -stopping time q with values in $[n, n + kn]$, we have $E|X_q|^2/a_n = O(1)$, uniformly in n and in q in the given class. The bounds also hold for the Y_i^n defined by (3.7).

Proof. The proof can be seen in the Appendix.

The following theorem gives us the largest window of averaging. It is largest in the sense that if $q_n = O(n)$, then the window is $O(n)$. The mutual independence of the increments of the Wiener process, which is the limit in the theorem, sheds additional light on the time scales which are used.

THEOREM 3.1. Under (1.2), (A2.1) and (A3.1)–(A3.5), $M^n(\cdot) \Rightarrow W(\cdot)$, a Wiener process with covariance Vt .

The proof will be divided into several parts. First we will show that it is sufficient to replace (1.1) by a simpler iteration. Then tightness of $\{M^n(\cdot)\}$ in the Skorohod topology is shown, and finally we prove that the limit of any weakly convergent subsequence is the asserted Wiener process.

Proof. It is notationally easier to do the proof if the rate at which $a_n \rightarrow 0$ is very slow. We will work with the additional assumption (see (1.2), (2.4))

$$(*) \quad \sup_{0 \leq i \leq kn} |\delta_{n,i}| \xrightarrow{n} 0$$

for each $k < \infty$. This will allow us to replace a_i by a_n when $0 \leq i - n \leq kn$. The modifications needed for the general case will be stated in Part 7.

Part 1. (In this part, (A3.1) can be replaced by $q_n a_n^2 \rightarrow 0$.) Define the quantity

$$\Pi(n, j) = \prod_{i=n}^j (I + a_i G), \quad j \geq n, \quad \Pi(n, n-1) = I.$$

Write (1.1) in the form

$$\begin{aligned} X_{n+1} &= X_n + a_n \bar{g}(X_n) + a_n \psi_n(X_n) \\ (3.3) \quad &= X_n + a_n G X_n + a_n \delta g(X_n) + a_n \psi_n(X_n). \end{aligned}$$

Then

$$\begin{aligned} (3.4) \quad X_{n+m+1} &= \Pi(n, n+m) X_n + \sum_{j=n}^{n+m} \Pi(j+1, n+m) a_j \delta g(X_j) \\ &\quad + \sum_{j=n}^{n+m} \Pi(j+1, n+m) a_j \psi_j(X_j) \\ &\equiv Q_{n,n+m}^1 + Q_{n,n+m}^2 + Y_{m+1}^n \end{aligned}$$

where the $Q_{n,j}^i$ and Y_m^n are defined in the obvious way.

It will be shown first that

$$(3.5) \quad \bar{Q}_n^i = \frac{1}{\sqrt{q_n}} \sum_{j=n}^{n+q_n t} E|Q_{n,j}^i| \xrightarrow{n} 0, \quad i = 1, 2.$$

This will allow us to drop the Q^1 and Q^2 terms in (3.4). The stability of G implies that there are $\lambda > 0$ and $c < \infty$ such that

$$(3.6) \quad \|\Pi(n, j)\| \leq ce^{-\lambda(t_j - t_n)}, \quad j \geq n.$$

Thus

$$E|Q_{n,j}^1| \leq ce^{-\lambda(t_j - t_n)} E|X_n|.$$

Using this bound and the estimate $E|X_n| = O(a_n^{1/2})$ from Lemma 3.1 yields

$$\bar{Q}_n^1 \leq O(1) \sum_{j=n}^{n+q_n t} e^{-\lambda(t_j - t_n)} O(\sqrt{a_n}) / \sqrt{q_n}.$$

By (A3.1), $1/\sqrt{a_n q_n} \rightarrow 0$. Using this and (2.3), (2.4) yields that

$$\bar{Q}_n^1 = O(1) \int_0^\infty e^{-\lambda s} ds / \sqrt{q_n a_n} \rightarrow 0.$$

Next we evaluate $Q_{n,n+m}^2$. For $m \in [n, kn]$, by Lemma 3.1 we have $E|\delta g(X_m)| = O(a_n)$. This and (3.6) yield

$$\begin{aligned} E|Q_{n,n+m}^2| &\leq O(1) \sum_{j=n}^{n+m} e^{-\lambda(t_{n+m} - t_{j+1})} a_j E|\delta g(X_j)| \\ &= O(1) a_n \int_0^\infty e^{-\lambda s} ds, \end{aligned}$$

which yields $E|Q_{n,n+m}^2| = O(a_n)$, $m \leq kn$. This implies that $\bar{Q}_n^2 \leq \sqrt{q_n t} O(a_n)$, which goes to zero as $n \rightarrow \infty$ by (A3.1).

Thus, to prove the theorem we can replace $\{X_m, m \geq n\}$ by the $\{Y_m^n, m \geq n\}$ process, which can be defined by

$$(3.7) \quad Y_{m+1}^n = (I + a_m G) Y_m^n + a_m \psi_m(X_m), m \geq n,$$

where we define $Y_n^n = 0$. Note that the stability of G in (A3.3) and the boundedness of $\{\psi_m(X_m)\}$ imply that $(Y_m^n, m \geq n)$ is bounded uniformly in n .

Part 2. Let $k > 0$ and let q be an \mathcal{F}_m -stopping time with values in $[n, n + kn]$. We next prove that

$$(3.8) \quad E|Y_q^n (E_q Y_j^n)'| = O(a_n^{3/2}) + O(a_n) E e^{-\lambda(t_j - t_q)},$$

for $kn \geq j \geq q$, where j and q are integers. A perturbed test function method will be used. For $n + kn \geq j \geq n$, define the ‘‘perturbations’’

$$(3.9) \quad \delta Y_j^n = \sum_{i=j}^\infty a_i E_j \psi_i(X_j) = O(a_j)$$

$$\tilde{Y}_j^n = Y_j^n + \delta Y_j^n$$

where the $O(a_j)$ value is due to (3.2). Note that the argument of the $\psi_i(\cdot)$ in (3.9) is X_j , the state at the lower index of summation.

Note the following:

$$(3.10a) \quad E_j Y_{j+1}^n - Y_j^n = a_j G Y_j^n + a_j E_j \psi_j(X_j),$$

$$(3.10b) \quad \begin{aligned} E_j \delta Y_{j+1}^n - \delta Y_j^n &= -a_j E_j \psi_j(X_j) \\ &+ \sum_{i=j+1}^{\infty} a_i E_j [\psi_i(X_{j+1}) - \psi_i(X_j)] \\ &= -a_j E_j \psi_j(X_j) + S_j, \end{aligned}$$

where S_j is defined in the obvious way. By (A3.4),

$$(3.11) \quad |S_j| = O(a_j^2)[|X_j| + 1].$$

Combining (3.10a), (3.10b) yields

$$(3.12) \quad E_j \tilde{Y}_{j+1}^n = (I + a_j G) \tilde{Y}_j^n + S_j - a_j G \delta Y_j^n.$$

Solving (3.12) yields

$$E_q \tilde{Y}_j^n = \Pi(q, j-1) \tilde{Y}_q^n + \sum_{i=q}^{j-1} \Pi(i, j-1) [E_q S_i - a_i E_q G \delta Y_i^n].$$

Hence, with the estimate $\delta Y_j^n = O(a_j)$ and (3.11), we can write

$$(3.13) \quad \begin{aligned} |E_q Y_j^n| &= O(1) e^{-\lambda(t_j - t_q)} |Y_q^n| \\ &+ O(1) \sum_{i=q}^{j-1} \Pi(i, j-1) a_i^2 (E_q |X_i| + 1) + O(a_n) \\ &= O(1) e^{-\lambda(t_j - t_q)} |Y_q^n| + O(a_n). \end{aligned}$$

By Lemma 3.1,

$$(3.14) \quad E|Y_q^n|^2 = O(a_n).$$

Now, combining (3.13) and (3.14) yields (3.8). Equation (3.13) will be used frequently in the sequel.

Part 3. Define

$$F^n(t) = \frac{1}{\sqrt{q_n}} \sum_{i=n}^{n+q_n t} Y_i^n.$$

By the results in Part 1, it is sufficient to prove the theorem for $F^n(\cdot)$ replacing $M^n(\cdot)$.

Tightness of $\{F^n(\cdot)\}$. Let $k < \infty$. Let $r(n)$ be a \mathcal{F}_m -stopping time, with values in $[n, n + kq_n]$. To prove tightness, it is sufficient ([7, Thm. 8.6] or, equivalently, [12, Thm. 3.3]) if $\sup_n E|F^n(t)| < \infty$ for each $t > 0$ and

$$(3.15) \quad \lim_{\delta \rightarrow 0} \limsup_n \sup_{r(n)} E|F^n(t_{r(n)} + \delta - t_n) - F^n(t_{r(n)} - t_n)|^2 = 0.$$

For notational simplicity, let the X_n and Y_j^n be real valued henceforth in the proof. The proof for the general case is the same. We can write

$$(3.16) \quad E|F^n(t_{r(n)} + \delta - t_n) - F^n(t_{r(n)} - t_n)|^2 = \frac{1}{q_n} E \sum_{i,j=r(n)}^{r(n)+q_n\delta} Y_i^n Y_j^n.$$

By (3.8) and the fact that $\sum_{j=m}^{\infty} e^{-\lambda(t_j - t_m)} a_j = O(1)$ uniformly in m , the above expression equals

$$\frac{1}{q_n} [\delta^2 q_n^2 O(a_n^{3/2}) + \delta q_n O(1)]$$

which goes to zero as needed due to (A3.1).

Part 4. We next show that the limit of any weakly convergent subsequence of $\{F^n(\cdot)\}$ is a martingale. Let $f(\cdot)$ be any bounded and continuous function of its arguments. For any integer p , fix $s \geq 0, \tau \geq 0$, and let $s_i \leq s, i = 1, \dots, p$. Then

$$\begin{aligned} Ef(F^n(s_i), i \leq p)[F^n(s + \tau) - F^n(s)] \\ = Ef(F^n(s_i), i \leq p)E_{n+q_ns}[F^n(s + \tau) - F^n(s)], \end{aligned}$$

where E_{n+q_ns} is the expectation given all data up to iterate $n + q_ns$ or, equivalently, given all data which is used to calculate $F^n(u), u \leq s$.

We have

$$\begin{aligned} E|E_{n+q_ns}F^n(s + \tau) - F^n(s)| \\ = E \frac{1}{\sqrt{q_n}} \left| E_{n+q_ns} \sum_{i=n+q_ns}^{n+q_ns+q_n\tau} Y_i^n \right|. \end{aligned}$$

By (3.13) and Lemma 3.1, this expression equals

$$\begin{aligned} (3.17) \quad & \frac{1}{\sqrt{q_n}} \left[q_n O(a_n) \tau + O(a_n^{1/2}) \sum_{i=n+q_ns}^{n+q_ns+q_n\tau} e^{-\lambda(t_i - t_n)} \right] \\ & \leq \sqrt{q_n} O(a_n) \tau + O(1) \sum_{i=n+q_ns}^{\infty} e^{-\lambda(t_i - t_n)} a_i / \sqrt{a_n q_n}, \end{aligned}$$

which goes to zero as $n \rightarrow \infty$ uniformly in any bounded τ -interval. Let $F(\cdot)$ denote the limit of a weakly convergent subsequence of $\{F^n(\cdot)\}$. By the fact that the right side of (3.16) is $O(\delta)$, $\{F^n(t)\}$ is uniformly integrable, for each $t < \infty$. This and the fact that expression (3.17) goes to zero as $n \rightarrow \infty$ yields

$$Ef(F(s_i), i \leq p)[F(s + \tau) - F(s)] = 0$$

for all s, τ, s_i, p and $f(\cdot)$ in the chosen classes. This implies that $F(\cdot)$ is a martingale. Since the discontinuities in $F^n(\cdot)$ are $O(q_n^{-1/2})$ and tend to zero as $n \rightarrow \infty$, $F(\cdot)$ is continuous. To prove that it is the asserted Wiener process, we need only identify its quadratic variation. This will be done in Part 6. In preparation for that, we need the following uniform integrability result.

Part 5. A uniform integrability result. For $s > 0, \tau > 0$, and given $T < \infty$ define the sets

$$I_\nu^n = [n + q_n s + \nu T/a_n, n + q_n s + (\nu + 1)T/a_n),$$

$\nu = 0, 1, \dots, \tau K_n - 1$, (assuming τK_n is an integer without loss of generality) where K_n is defined by

$$(3.18) \quad K_n(T/a_n) = q_n.$$

Then, starting at time s to “cover” $[s, s + \tau)$, we need $K_n \tau$ groups of (T/a_n) iterates each. Define

$$\delta F_\nu^n = \left(\frac{a_n}{T}\right)^{1/2} \sum_{i \in I_\nu^n} Y_i^n.$$

We will show that, for each T ,

$$(3.19) \quad \sup_{\nu \leq K_n \tau} E|\delta F_\nu^n|^4 < \infty.$$

For simplicity of notation, we work only with the real valued Y_i^n case. For $\nu \leq K_n \tau$, there is a $k < \infty$, such that the indices i for which Y_i^n is in δF_ν^n are all in the interval $[n, n + kn]$, so that (2.4) and the ratio $a_n/a_i \approx 1$ can be used. We have

$$E|\delta F_\nu^n|^4 = O(1) \left(\frac{a_n}{T}\right)^2 \sum_{i \leq j \leq k \leq \ell} EY_i^n Y_j^n Y_k^n Y_\ell^n,$$

where the indices vary over the set I_ν^n , subject to the indicated inequalities. By (3.13), this expression can be written as

$$\begin{aligned} & O(1) \left(\frac{a_n}{T}\right)^2 \sum_{i \leq j \leq k \leq \ell} EY_i^n Y_j^n Y_k^n E_k Y_\ell^n \\ & \leq O(1) \left(\frac{a_n}{T}\right)^2 \sum_{i \leq j \leq k \leq \ell} E|Y_i^n Y_j^n Y_k^n| \\ & \quad \times [e^{-\lambda(t_\ell - t_k)} |Y_k^n| + a_n]. \end{aligned}$$

Now, using Lemma 3.1 and the fact that $\sum_{j \geq n} e^{-\lambda(t_j - t_n)} a_j = O(1)$, uniformly in n yields that the sum is $O(1)$, uniformly in n and in the range of ν in question, which proves the assertion (3.19).

Part 6. The quadratic variation of the limit process. For notational simplicity, we continue to work with the case of real valued X_n, Y_i^n . We first present some consequences of the weak convergence in Theorem 2.1. Recall that

$$\sqrt{\frac{a_n}{T}} \sum_{i \in I_\nu^n} X_i \rightarrow N(0, V_T),$$

where the symbol $(\rightarrow N(0, V_T))$ means convergence in distribution to a normally distributed random variable with mean zero and variance V_T . Recall that $V_T = V + O(1/t)$.

Also, by the weak convergence in Theorem 2.1, the set

$$\left\{ \left(\frac{a_n}{T} \right)^{1/2} \sum_{i \in I_{k_j}^n} X_i, \quad j = 1, \dots, q \right\}$$

converges in distribution to a set of normal random variables $\{Z_1, \dots, Z_q\}$, with mean zero and covariance

$$\text{cov}[Z_i, Z_j] = O\left(e^{-\lambda T|k_i - k_j|}\right).$$

Furthermore, there is $\varepsilon(T) \rightarrow 0$ as $T \rightarrow \infty$ such that the following holds if $N_n \rightarrow \infty$ slowly enough:

$$(3.20) \quad \frac{1}{N_n} \sum_{\nu=1}^{N_n} (\delta F_\nu^n)^2 \xrightarrow{P} V(T),$$

where $V(T) \in [V - \varepsilon(T), V + \varepsilon(T)]$ and \xrightarrow{P} denotes convergence in probability. The result is a consequence of the weak convergence in Theorem 2.1, the equivalences and estimates in Part 1 of the proof, the uniform integrability of $\{(\delta F_\nu^n)^2, \nu \leq kK_n, n\}$ for any $k < \infty$, and a law of large numbers. By the uniform integrability, (3.20) also holds in the mean.

The bound on ν is chosen to assure that the time indices i of all the iterates Y_i^n fall in the range $[n, n + 2kn]$, so that (2.4) can be used.

Let n index a weakly convergent subsequence of $\{F^n(\cdot)\}$ with limit $F(\cdot)$. Let $f(\cdot)$ be a bounded continuous function, and let $\phi(\cdot)$ be continuous with compact support with the first two derivatives being continuous. For any integer p , let $s_i \leq s, i \leq p$, and let $\tau > 0$. To get the quadratic variation result, we need to show that for all such $f(\cdot), \phi(\cdot), s_i, \tau, s, p$,

$$\begin{aligned} & Ef(F^n(s_i), i \leq p)[\phi(F^n(s + \tau)) - \phi(F^n(s))] \\ (3.21) \quad & \rightarrow Ef(F(s_i), i \leq p)[\phi(F(s + \tau)) - \phi(F(s))] \\ & = Ef(F(s_i), i \leq p) \left[\frac{V}{2} \int_s^{s+\tau} \phi_{FF}(F(u)) du \right]. \end{aligned}$$

Note that

$$F^n\left(s + \frac{(\nu + 1)T}{a_n q_n}\right) - F^n\left(s + \frac{\nu T}{a_n q_n}\right) = \frac{1}{K_n^{1/2}} \delta F_\nu^n.$$

Now, expanding the left side of (3.21) yields

$$\begin{aligned} (3.22) \quad & Ef(F^n(s_i), i \leq p) \\ & \times \left[\frac{1}{K_n^{1/2}} \sum_{\nu=1}^{K_n \tau} \phi_F\left(F^n\left(s + \frac{\nu T}{a_n q_n}\right)\right) \delta F_\nu^n + \frac{1}{2K_n} \sum_{\nu=1}^{K_n \tau} \phi_{FF}\left(F^n\left(s + \frac{\nu T}{a_n q_n}\right)\right) (\delta F_\nu^n)^2 \right. \\ & \left. + \frac{1}{K_n^{3/2}} \sum_{\nu=1}^{K_n \tau} O(|\delta F_\nu^n|^3) \right]. \end{aligned}$$

By the results in Parts 1, 4, and 5, (3.22) is asymptotically equivalent to

$$(3.23) \quad Ef(F^n(s_i), i \leq p) \left[\frac{1}{2K_n} \sum_{\nu=1}^{K_n \tau} \phi_{FF} \left(F^n \left(s + \frac{\nu T}{a_n q_n} \right) \right) (\delta F_\nu^n)^2 \right].$$

By the tightness of $\{F^n(\cdot)\}$ and the uniform integrability of $\{|\delta F_\nu^n|^2\}$ shown by (3.19), we can “delay” the time argument of the $F^n(\cdot)$ in the ϕ_{FF} in (3.23) by an amount which goes to zero as $n \rightarrow \infty$, without changing the asymptotic value. This observation allows us to regroup the summands in (3.23) so that (3.20) can be used. In fact, (3.23) is asymptotically equivalent to (3.24), where we define v_n by $\tau K_n = N_n v_n$, with $N_n \rightarrow \infty$ as slowly as we wish but such that $K_n \rightarrow \infty$ and $v_n \rightarrow \infty$:

$$(3.24) \quad Ef(F^n(s_i), i \leq p) \times \left[\frac{1}{2v_n} \sum_{\nu=1}^{v_n} \left\{ \phi_{FF} \left(F^n \left(s + \frac{\nu N_n T}{a_n q_n} \right) \right) \frac{1}{N_n} \sum_{u=\nu N_n}^{\nu N_n + N_n - 1} (\delta F_u^n)^2 \right\} \right].$$

Finally, applying (3.20) to (3.24), taking limits as $n \rightarrow \infty$, and using the arbitrariness of T and $\varepsilon(T)$ yields the desired result, namely, the right side of (3.21).

Part 7. Dropping condition ().* When (*) is dropped, we need to regroup certain terms so that the same asymptotic expansions will hold. Define the increasing sequence ρ_v (depending on n) recursively by $\rho_0 = 0$, and for $v \geq 1$

$$\sum_{n+q_n s + \rho_{v-1}}^{n+q_n s + \rho_v} a_j \rightarrow T$$

as $n \rightarrow \infty$. Redefine the sets of indices I_v^n to be $I_v^n = [n+q_n s + \rho_{v-1}, n+q_n s + \rho_v)$. Thus the sums of the a_i in each set equal T asymptotically. Set $m(n, v) = n + q_n s + \rho_{v-1}$, the first index in the set I_v^n . In sums of the form $(a_n/T)^{1/2} \sum I_v^n$, replace the a_n by $a_{m(n,v)}$. Define $J_n = \min\{\alpha : \rho_\alpha \geq q_n \tau\}$. The J_n replace the K_n in the proof. Finally in the expansions from (3.22) to (3.24), replace the vT/a_n by ρ_v .

With these changes the proof goes through as done above. \square

Appendix.

Proof of Lemma 3.1. The proof will be given for the $\{X_n\}$ only. The proof for the $\{Y_i^n\}$ follows from this, and the details are omitted.

Part 1. Mean square bounds. A perturbed Liapunov function method will be used. Define the perturbation

$$V_1(x, n) = \sum_{j=n}^{\infty} a_j V'_x(x) E_n \psi_j(x) = O(a_n) |V_x(x)|,$$

where the right-hand inequality is due to (A3.2). We can write

$$(1) \quad \begin{aligned} E_n V(X_{n+1}) - V(X_n) &= a_n V'_x(X_n) \bar{g}(X_n) \\ &\quad + a_n V'_x(X_n) \psi_n(X_n) + a_n^2 O(1) |g(X_n, \xi_n)|^2 \end{aligned}$$

$$(2) \quad \begin{aligned} E_n V_1(X_{n+1}, n+1) - V_1(X_n, n) \\ = -a_n V'_x(X_n) \psi_n(X_n) + \sum_{j=n+1}^{\infty} E_n a_j [V'_x(X_{n+1}) \psi_j(X_{n+1}) - V'_x(X_n) \psi_j(X_n)]. \end{aligned}$$

The fact that the second term on the right side of (1) is the negative of the first term on the right side of (2) is the essential motivation behind the construction of $V_1(\cdot)$. Rewriting the last term on the right side of (2) as the sum of the left-hand sides in (3a), (3b) following and bounding them by use of (A3.2)–(A3.5) yields:

$$(3a) \quad \left| \sum_{n+1}^{\infty} E_n a_j V'_x(X_n) (\psi_j(X_{n+1}) - \psi_j(X_n)) \right| \\ \leq |V_x(X_n)| O(a_n^2) [V^{1/2}(X_n) + 1] \leq O(a_n^2) [V(X_n) + 1],$$

$$(3b) \quad \left| \sum_{n+1}^{\infty} a_j E_n [V'_x(X_{n+1}) - V'_x(X_n)] \psi_j(X_{n+1}) \right| \\ \leq O(a_n^2) |g(X_n, \xi_n)| \leq O(a_n^2) [V(X_n) + 1].$$

Define the perturbed Liapunov function $\tilde{V}_n = V(X_n) + V_1(X_n, n)$. Putting the estimates (2) and (3) together yields

$$E_n \tilde{V}_{n+1} - \tilde{V}_n \leq -a_n \gamma V(X_n) + O(a_n^2) \\ + O(a_n^2) [1 + V(X_n)]$$

and

$$(4) \quad E_n \tilde{V}_{n+1} - \tilde{V}_n \leq -\frac{a_n}{2} \gamma \tilde{V}_n + O(a_n^2).$$

Equation (4) implies that $\{E\tilde{V}_n/a_n, n < \infty\}$ is bounded from above. Then, using this and the estimate $V_1(x, n) = O(a_n)[V(x) + 1]$ yields the boundedness of $\{EV(X_n)/a_n, n < \infty\}$. This latter bound and the first equation of (A3.5) yield the boundedness of $\{E|X_n|^2/a_n, n < \infty\}$.

Equation (4) implies that there are β_n such that $E|\beta_n| < \infty$, $E_n \beta_n = 0$, and

$$(5) \quad \tilde{V}_{n+1} - \tilde{V}_n \leq \frac{-a_n}{2} \gamma \tilde{V}_n + O(a_n^2) + \beta_n,$$

from which we get

$$E_n \tilde{V}_q = O(a_n)$$

for any \mathcal{F}_m -stopping time q with values in $[n, n + kn]$. This, together with the above given bound on $V_1(x, n)$, yields the second assertion of the lemma.

Part 2. Fourth moments. We now prove $E|X_n|^4 = O(a_n^2)$. The procedure will be similar to that used in Part 1. Define the perturbation

$$(6) \quad V_2(x, n) = 2 \sum_{j=n}^{\infty} a_j E_n V(x) V'_x(x) \psi_j(x) = O(a_n) |V(x) V_x(x)|,$$

and the perturbed Liapunov function

$$(7) \quad \hat{V}(x, n) = V^2(x) + V_2(x, n).$$

We use \tilde{X}_n and \hat{X}_n to denote vectors in the interval $[X_n, X_{n+1}]$, and their values might change from case to case. Proceeding as for the second-order case, by a truncated Taylor series expansion we can write

$$(8) \quad E_n V^2(X_{n+1}) - V^2(X_n) = B_1 + B_2 + B_3 + B_4,$$

where

$$\begin{aligned} B_1 &= 2a_n V(X_n) V'_x(X_n) \bar{g}(X_n), \\ B_2 &= 2a_n E_n V(X_n) V'_x(X_n) \psi_n(X_n), \\ B_3 &= a_n^2 E_n g'(X_n, \xi_n) V_x(\tilde{X}_n) V'_x(\tilde{X}_n) g(X_n, \xi_n), \\ B_4 &= a_n^2 E_n V(\tilde{X}_n) g'(X_n, \xi_n) V_{xx}(\tilde{X}_n) g(X_n, \xi_n). \end{aligned}$$

Furthermore,

$$(9) \quad E_n V_2(X_{n+1}, n+1) - V_2(X_n, n) = B_5 + B_6,$$

where

$$\begin{aligned} B_5 &= -2a_n E_n V(X_n) V'_x(X_n) \psi_n(X_n), \\ B_6 &= 2 \sum_{j=n+1}^{\infty} a_j E_n \left[V(X_{n+1}) V'_x(X_{n+1}) \psi_j(X_{n+1}) - V(X_n) V'_x(X_n) \psi_j(X_n) \right]. \end{aligned}$$

The terms $B_i, i = 1, \dots, 6$, will now be bounded. Heavy use will be made of the inequalities in (A3.5) and the fact that $EV(X_n) = O(a_n)$ by Part 1 of the proof. We have

$$B_1 \leq -2a_n \gamma V^2(X_n).$$

B_2 is cancelled by B_5 . For appropriate \tilde{X}_n and \hat{X}_n ,

$$\begin{aligned} B_3 &= a_n^2 E_n g'(X_n, \xi_n) [V_x(X_n) + V_{xx}(\hat{X}_n)(\tilde{X}_n - X_n)] \\ &\quad \times [V_x(X_n) + V_{xx}(\hat{X}_n)(\tilde{X}_n - X_n)]' g(X_n, \xi_n) \\ &= C_1 + C_2 + C_3, \end{aligned}$$

where

$$\begin{aligned} EC_1 &= 2a_n^2 E g'(X_n, \xi_n) V_x(X_n) V'_x(X_n) g(X_n, \xi_n) \\ EC_2 &= O(a_n^2) E |g(X_n, \xi_n)|^2 |V_x(X_n)| |\tilde{X}_n - X_n| \\ EC_3 &= O(a_n^2) E |g(X_n, \xi_n)|^2 |\tilde{X}_n - X_n|^2. \end{aligned}$$

We have the following bounds:

$$\begin{aligned}
 EC_1 &\leq O(a_n^2)EV(X_n)(V(X_n) + 1) \\
 &\leq O(a_n^2)EV^2(X_n) + O(a_n^3) \\
 EC_2 &\leq O(a_n^2)E(V(X_n) + 1)|V_x(X_n)||\tilde{X}_n - X_n| \\
 &\leq O(a_n^2)E(V(X_n) + 1)|V_x(X_n)||X_{n+1} - X_n| \\
 &\leq O(a_n^3)E(V(X_n) + 1)^2 \leq O(a_n^3)EV^2(X_n) + O(a_n^3), \\
 EC_3 &\leq O(a_n^2)E(V(X_n) + 1)a_n^2(V(X_n) + 1) \\
 &\leq O(a_n^4)EV^2(X_n) + O(a_n^4).
 \end{aligned}$$

By a similar method, the other terms can be shown to satisfy

$$EB_i = O(a_n^2)EV^2(X_n) + O(a_n^3), \quad i = 4, 5, 6.$$

Finally, putting these estimates together yields

$$\begin{aligned}
 (10) \quad E\hat{V}(X_{n+1}, n+1) - E\hat{V}(X_n) &\leq -\gamma a_n EV^2(X_n) \\
 &\quad + O(a_n^2)EV^2(X_n) + O(a_n^3).
 \end{aligned}$$

Note that

$$\begin{aligned}
 (11) \quad |V_2(X_n, n)| &= O(a_n)V(X_n)(V(X_n) + 1) \\
 &\leq O(a_n)EV^2(X_n) + O(a_n^2).
 \end{aligned}$$

Using (10) and (11) yields, for large n ,

$$(12) \quad E\hat{V}(X_{n+1}, n+1) - E\hat{V}(X_n, n) \leq -\frac{\gamma}{2}a_n E\hat{V}(X_n, n) + O(a_n^3).$$

Then, following the procedure of Part 1, we get

$$\sup_n EV^2(X_n)/a_n^2 < \infty$$

and

$$E|X_n|^4/a_n^2 < \infty. \quad \square$$

REFERENCES

- [1] B. T. POLYAK, *New stochastic approximation type procedures*, Automat. i Telemekh., 7 (1990), pp. 98–107.
- [2] B. T. POLYAK AND A. B. JUDITSKY, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.
- [3] D. RUPPERT, *Efficient estimators from a slowly convergent Robbins–Monro process*, Tech. Report, No. 781, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988.
- [4] G. YIN, *Stochastic approximation via averaging; Polyak's approach revisited*, Lecture Notes in Economics and Mathematical Systems 374, G. Pflug and U. Dieter, eds. Springer-Verlag, Berlin, 1992, pp. 119–134.

- [5] G. YIN, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics, 36 (1992), pp. 245–264.
- [6] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [7] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
- [8] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Applied Math. Sciences Series, Springer-Verlag, Berlin, New York, 1978.
- [9] A. BENVENISTE, M. METIVIER, AND P. PRIORET, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, New York, 1990. (Translated from the French.)
- [10] H. J. KUSHNER, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhauser, Boston, 1990.
- [11] H. J. KUSHNER AND HAI HUANG, *Averaging methods for the asymptotic analysis of learning and adaptive systems with small adjustment rate*, SIAM J. Control Optim., 19 (1981), pp. 635–650.
- [12] ———, *Approximation and Weak Convergence Methods for Stochastic Processes with Applications to Stochastic Systems Theory*, M.I.T. Press, Cambridge, 1984.
- [13] H. J. KUSHNER AND G. YIN, *Asymptotic properties of distributed and communicating stochastic approximation algorithms*, SIAM J. Control Optim., 25 (1987), pp. 1266–1290.
- [14] H. J. KUSHNER AND A. SHWARTZ, *An invariant measure approach to the convergence of stochastic approximations with state dependent noise*, SIAM J. Control Optim., 22 (1984), pp. 13–27.
- [15] H. J. KUSHNER, *Stochastic approximation with discontinuous dynamics and state dependent noise*, J. Math. Anal. Appl., 82 (1981), pp. 527–542.

IDENTIFIABLE SURFACES IN CONSTRAINED OPTIMIZATION*

STEPHEN J. WRIGHT†

Abstract. The concept of a “class- C^p identifiable surface” of a convex set in Euclidean space is introduced. The paper shows how the smoothness of these surfaces is related to the smoothness of the projection operator and presents finite identification results for certain algorithms for minimization of a function over this set. The work uses a partially geometric view of constrained optimization to generalize previous finite identification results.

Key words. constrained optimization, active set identification

AMS subject classifications. 90C25, 53A07, 26B10

1. Introduction. Here, we investigate the problem

$$(1) \quad \min_{x \in \Omega} F(x),$$

where F is continuously differentiable and $\Omega \subset \mathbb{R}^n$ is closed and convex. In particular, we are interested in finding subsets of Ω that can be identified by an optimization algorithm after a finite number of iterations. That is, if the solution x^* lies in one such subset, the iterates generated by the algorithm should eventually enter and remain within that subset. In the case in which Ω is defined by a set of algebraic inequalities, this property of the iterates corresponds to identifying the active constraints, and when Ω is a polyhedron, it means identifying the face, edge, or vertex, upon which the solution x^* lies.

The first-order conditions for x^* to be a solution of (1) are

$$-\nabla F(x^*) \in N(x^*),$$

where $N(x^*)$ is the normal cone to Ω at x^* . To prove the finite identification (capture) results, we assume a nondegeneracy condition due to Dunn [4]. This is stated simply as

$$(2) \quad -\nabla F(x^*) \in \text{ri}(N(x^*)),$$

where $\Lambda \subset \mathbb{R}^n$ and $\text{ri}(\Lambda)$ is the interior of Λ relative to $\text{aff}(\Lambda)$, the affine hull of Λ . A condition equivalent to (2) was assumed by Gafni and Bertsekas [7] for the case of polyhedral sets. The condition (2), which is a geometric generalization of the strict complementarity condition of nonlinear programming, has been used in the convergence analysis of Dunn [4] and Burke and Moré [1]. Both these papers specify similar classes of subsets of Ω that are finitely identifiable by gradient projection and Newton-like algorithms. We define these *open facets* as in [4].

DEFINITION 1.

(a) For any closed convex cone $K \subset \mathbb{R}^n$, we use K° to denote the polar of K , and define the lineality $\text{lin}(K)$ to be $(K^\circ)^\perp$;

(b) Let $T(x)$ be the tangent cone to Ω at x as defined in Clarke [2, Thm. 2.4.5]; the normal cone is $N(x) = T(x)^\circ$. A nonempty subset $S \subset \Omega$ is an *open facet* if the

* Received by the editors October 7, 1991; accepted for publication (in revised form) April 9, 1992. This research was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U. S. Department of Energy, under contract W-31-109-Eng-38.

† Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439.

set $V = x + \text{lin}(T(x))$ is independent of $x \in S$, and $S = \text{int}_V(\Omega \cap V)$, where $\text{int}_V(\cdot)$ denotes interior with respect to V .

It is easy to show that open facets are convex. When Ω is polyhedral, it can be partitioned into open facets, but when Ω has some curved boundaries, this is not the case. As an example, consider the set defined in [1, Eq. (2.2)]:

$$\Omega_1 = \left\{ (\xi_1, \xi_2) \mid \xi_2 \leq \sqrt{1 - \xi_1^2}, 0 \leq \xi_1 \leq 1 \right\}.$$

The open facets in this set are its interior, the point $(0, 1)$ and the edges $\{(0, \xi_2) \mid \xi_2 < 1\}$ and $\{(1, \xi_2) \mid \xi_2 < 0\}$. No subset of the curved face $\{(\xi_1, \xi_2) \mid \xi_1^2 + \xi_2^2 = 1, 0 < \xi_1 \leq 1\}$ satisfies Definition 1.

When Ω is defined by algebraic inequalities, that is,

$$(3) \quad \Omega = \{\xi \mid g_i(\xi) \leq 0, i = 1, \dots, m\},$$

it is often assumed that the g_i are C^2 and that the set

$$(4) \quad \{\nabla g_i(x) \mid i \in \mathcal{A}(x)\}, \quad \text{where} \quad \mathcal{A}(x) = \{i \mid 1 \leq i \leq m, g_i(x) = 0\}$$

is linearly independent. In this case, the nondegeneracy condition (2) (which reduces to the standard strict complementarity condition) ensures that surfaces defined by a particular active index set $\mathcal{A} \subset \{1, 2, \dots, m\}$ are finitely identifiable by a number of standard algorithms. Note that Ω_1 is not definable in the form (3),(4) for $g_i \in C^2$, since there is a curvature discontinuity in the boundary at $(1, 0)$. If we allow g_i to be only C^1 , then Ω_1 is definable as (3),(4), but then the curved surface is indistinguishable from the face $\{(1, \xi_2) \mid \xi_2 \leq 0\}$.

The next section defines the concept of a “class- C^p identifiable surface.” Loosely speaking, such a surface S is usually a connected “patch” on $\partial\Omega$, which is locally parametrizable by a collection of C^p functions, for some integer $p \geq 1$. (The interior of Ω is defined to be a class- C^∞ surface.) Moreover, these functions can be defined so that their gradients can enclose any given ray in the relative interior of $N(x)$, where x is a given point in S . We prove that open facets and subsets of (3) that are defined by particular choices of \mathcal{A} are identifiable surfaces. (For the set Ω_1 , the curved boundary, with its two endpoints excluded, is also an identifiable surface.) We show that class- C^p identifiable surfaces generate connected open regions in the exterior of Ω , within which the operation of projection onto Ω is $p - 1$ times continuously differentiable. In §3, we prove finite identification results for gradient projection and Newton-like algorithms.

In the remainder of the paper, $\|\cdot\|$ denotes the Euclidean norm, B denotes the open unit ball $\{\xi \in \mathbb{R}^n \mid \|\xi\| < 1\}$, and $\text{co}\{\cdot\}$ denotes the convex hull of a set of vectors. The projection operator and distance function are defined as follows, with reference to any closed subset A of \mathbb{R}^n :

$$P_A(y) = \min_{\bar{y} \in A} \frac{1}{2} \|\bar{y} - y\|, \quad d_A(y) = \arg \min_{\bar{y} \in A} \frac{1}{2} \|\bar{y} - y\|.$$

We use $P(\cdot)$ as shorthand for $P_\Omega(\cdot)$. Given a collection of functions $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, r$, we frequently use the notation

$$g(x) = [g_1(x), g_2(x), \dots, g_r(x)]^T, \quad \nabla g(x) = [\nabla g_1(x) \mid \dots \mid \nabla g_r(x)].$$

2. Identifiable surfaces and smoothness of the projection operator.

Throughout the remainder of this paper, we make this following assumption.

Assumption 1. Ω is closed and convex and has an interior in \mathbb{R}^n .

The last part of this assumption is made for convenience. If it does not hold, the results of this section can be recovered by restricting attention to $\text{aff}(\Omega)$.

DEFINITION 2. A connected set $S \subset \Omega$ is a *class- C^p identifiable surface*, p a positive integer, if either

- (a) S is an open subset of $\text{int}(\Omega)$, or
- (b) $S \subset \partial\Omega$, and for any $y \in \mathbb{R}^n \setminus \Omega$ such that $\bar{y} = P(y) \in S$ and $y - \bar{y} \in \text{ri}(N(\bar{y}))$, there exist functions g_i , $i = 1, \dots, r = r(y)$, and a constant $\epsilon = \epsilon(y) > 0$, such that
 - (i) $g_i \in C^p(\bar{y} + \epsilon B)$, $i = 1, \dots, r$;
 - (ii) $\{\nabla g_i(\bar{z}), i = 1, \dots, r\}$ is linearly independent for all $\bar{z} \in S(\bar{y}; \epsilon) \triangleq (\bar{y} + \epsilon B) \cap S$;
 - (iii) $\text{co}\{\nabla g_i(\bar{z}), i = 1, \dots, r\} \subset N(\bar{z})$ for all $\bar{z} \in S(\bar{y}; \epsilon)$;
 - (iv) $y - \bar{y} \in \text{ri}(\text{co}\{\nabla g_i(\bar{y}), i = 1, \dots, r\})$; and
 - (v) $S(\bar{y}; \epsilon) = \{\bar{z} \mid \|\bar{z} - \bar{y}\| < \epsilon, g_i(\bar{z}) = 0, i = 1, \dots, r\}$.

Obviously, if S is a class- C^p identifiable surface, then it is also a class- C^q identifiable surface, for any q with $1 \leq q < p$.

There are some significant differences between the functions g_i that are used in Definition 2 and the ones used in the algebraic parametrization (3). First, we are now only seeking a *local* parametrization. Second, and more importantly, we are trying to parametrize the surface S and a piece of the normal cone to Ω at points on S , rather than the set Ω itself. This latter property allows a wider range of sets to be decomposed into an intuitively reasonable collection of surfaces than was possible in the case of open facets or (3).

Before proceeding, to give a sense of how this definition differs from that of open facets and from (3), we review the example Ω_1 from §1, and give two more examples. The interior of Ω_1 , the point $(0, 1)$ and the edge $\{(0, \xi_2) \mid \xi_2 < 1\}$ are class- C^∞ identifiable surfaces. The remaining surface defined by

$$\left\{ (\xi_1, \xi_2) \mid \xi_2 = \sqrt{1 - \xi_1^2}, 0 \leq \xi_1 < 1 \right\} \cup \{(1, \xi_2) \mid \xi_2 < 0\} \cup \{(1, 0)\}$$

is class- C^1 identifiable. Each of the first two component subsets is class- C^∞ identifiable.

A second example is an inverted cone in \mathbb{R}^3 , whose apex is at the origin:

$$\Omega_2 = \left\{ (\xi_1, \xi_2, \xi_3) \mid \xi_3 \geq \sqrt{\xi_1^2 + \xi_2^2}, 0 \leq \xi_3 \leq 1 \right\}.$$

Ω_2 has just three open facets: the point $(0, 0, 0)$, the circular face $\{(\xi_1, \xi_2, 1) \mid \xi_1^2 + \xi_2^2 < 1\}$, and the interior. A finite algebraic parametrization (3),(4) is apparently not possible, even if we allow $g_i \in C^1$ (the difficulty is, of course, at the apex). However, the whole set *can* be partitioned into five maximal class- C^∞ identifiable surfaces. They are the three open facets just mentioned, the circle $\{(\xi_1, \xi_2, 1) \mid \xi_1^2 + \xi_2^2 = 1\}$, and the curved face $\{(\xi_1, \xi_2, \xi_3) \mid \xi_3 = \sqrt{\xi_1^2 + \xi_2^2}, 0 < \xi_3 < 1\}$. To show how the definition is satisfied in the case of $(0, 0, 0)$, take some $y \in \text{int}(N(0, 0, 0))$. Then $y = (y_1, y_2, y_3)$, with $y_3 < -\sqrt{y_1^2 + y_2^2}$. Clearly, we can choose $\gamma > 0$ such that

$y + \gamma B \subset \text{int}(N(0, 0, 0))$. Define three vectors as follows:

$$\begin{aligned} y^{(1)} &= (y_1 + \gamma, y_2, y_3) \\ y^{(2)} &= (y_1 - (1/2)\gamma, y_2 + (\sqrt{3}/2)\gamma, y_3) \\ y^{(3)} &= (y_1 - (1/2)\gamma, y_2 - (\sqrt{3}/2)\gamma, y_3). \end{aligned}$$

Elementary manipulation shows that these are linearly independent and that $y = (1/3)(y^{(1)} + y^{(2)} + y^{(3)})$. If we define $g_i(z) = z^T y^{(i)}$, the five conditions in Definition 2 are easily verified.

A third example is the set

$$\Omega_3 = \{(\xi_1, \xi_2, \xi_3) \mid \xi_3 \geq \xi_1^2 + |\xi_2| + \xi_2^{\sqrt{2}}\}.$$

This set is representable in the form (3),(4) by splitting the inequality into two (for the two possibilities $|\xi_2| = \pm \xi_2$), but the g_i are only C^1 . An algebraic parametrization that uses C^p functions, $p \geq 2$, is not possible. There are no open facets, except the interior. However, the set can be partitioned into four class- C^∞ identifiable surfaces. These are the interior, the face defined by

$$\{(\xi_1, \xi_2, \xi_3) \mid \xi_2 > 0, \xi_3 = \xi_1^2 + \xi_2 + \xi_2^{\sqrt{2}}\}$$

and its counterpart

$$\{(\xi_1, \xi_2, \xi_3) \mid \xi_2 < 0, \xi_3 = \xi_1^2 - \xi_2 + \xi_2^{\sqrt{2}}\},$$

and the ridge

$$\{(\xi_1, 0, \xi_3) \mid \xi_3 = \xi_1^2\}.$$

The ridge can be made to fit the definition by taking

$$\begin{aligned} g_1(\xi) &= \xi_1^2 + \xi_2 - \xi_3 \\ g_2(\xi) &= \xi_1^2 - \xi_2 - \xi_3, \end{aligned}$$

independently of the choice of $y \in \text{ri}(N(\xi))$.

The concept of a class- C^p identifiable surface is, in a certain sense, a generalization of the concept of a class- $C^{p,\alpha}$ boundary of a bounded domain $\Omega \subset \mathbb{R}^n$, as used extensively in the theory of partial differential equations (see, for example, the definition on page 94 of Gilbarg and Trudinger [9]). In fact, if Ω is convex, closed, and bounded, and its boundary $\partial\Omega$ is of class- $C^{p,0}$ according to the latter definition, then it can be partitioned into a class- C^∞ identifiable surface ($\text{int}(\Omega)$) and a class- C^p identifiable surface ($\partial\Omega$). Such sets have no “edges” or “corners”—the value of r corresponding to each $y \in \mathbb{R}^n \setminus \Omega$ is 1—and hence, they are not very interesting from the viewpoint of this paper.

We now derive some elementary properties of identifiable surfaces and the functions g_i that are used to describe them. We focus on the case $S \subset \partial\Omega$, since the corresponding results for $S \subset \text{int}(\Omega)$ are trivial.

LEMMA 2.1. *Let S be a class- C^p identifiable surface with $S \subset \partial\Omega$ and $p \geq 1$, and let $y \in \mathbb{R}^n \setminus \Omega$ be such that $\bar{y} = P(y) \in S$ and $y - \bar{y} \in \text{ri}(N(\bar{y}))$. Suppose that $r = r(y)$, $\epsilon = \epsilon(y)$, and g_i , $i = 1, \dots, r$ are chosen as in Definition 2. Then, for all $\bar{z} \in S(\bar{y}; \epsilon)$,*

- (i) $T_S(\bar{z}) \subset T(\bar{z})$, where $T_S(\cdot)$ is the tangent cone with respect to S , as defined in Clarke [2, Thm. 2.4.5];
- (ii) $T_S(\bar{z}) = \{s \mid s^T \nabla g_i(\bar{z}) = 0, i = 1, \dots, r\}$;
- (iii) $\text{lin}(T(\bar{z}))^\perp = \text{aff}(N(\bar{z})) = \text{span}\{\nabla g_i(\bar{z}), i = 1, \dots, r\} = T_S(\bar{z})^\perp$;
- (iv) $\text{ri}(\text{co}\{\nabla g_i(\bar{z}), i = 1, \dots, r\}) \subset \text{ri}(N(\bar{z}))$.
- (v) if $p \geq 2$, the projection of $\nabla^2 g_i(\bar{z})$, $i = 1, \dots, r$, onto $T_S(\bar{z})$ is positive semidefinite.

Proof.

(i) If $v \in T_S(\bar{z})$, it follows from the definition of tangent cone that for any sequence $\{t_j\}$ with $t_j \downarrow 0$ there is a sequence v_j such that $\bar{z} + t_j v_j \in S$ and $v_j \rightarrow v$. Since $S \subset \Omega$ and $\bar{z} + t_j v_j \in S$,

$$0 \leq d_\Omega(\bar{z} + t_j v) \leq d_S(\bar{z} + t_j v) = d_S(\bar{z} + t_j v_j + t_j(v - v_j)) \leq t_j \|v - v_j\|.$$

Hence,

$$0 \leq \lim_{j \rightarrow \infty} \frac{d_\Omega(\bar{z} + t_j v) - d_\Omega(\bar{z})}{t_j} \leq \lim_{j \rightarrow \infty} \|v - v_j\| = 0.$$

Since t_j is an arbitrary decreasing sequence,

$$d'_\Omega(\bar{z}; v) = \lim_{t \downarrow 0} \frac{d_\Omega(\bar{z} + tv) - d_\Omega(\bar{z})}{t} = 0,$$

and so, by a result of Clarke [2, p. 53], $v \in T(\bar{z})$.

(ii) This is a standard result, which follows from Definition 2(v).

(iii) We prove the second equality. By Definition 2(iii), $\text{span}\{\nabla g_i(\bar{z}), i = 1, \dots, r\} \subset \text{aff}(N(\bar{z}))$. Since both sets are subspaces, the containment can be strict only if there is some $v \in \text{aff}(N(\bar{z}))$ with $v \neq 0$ such that $v^T \nabla g_i(\bar{z}) = 0$, $i = 1, \dots, r$, that is, $v \in T_S(\bar{z})$. Clearly, also, $-v \in T_S(\bar{z})$. Part (i) of this Lemma implies that v and $-v$ are in $T(\bar{z})$, and hence, $v \in \text{lin}(T(\bar{z})) = N(\bar{z})^\perp$. Hence, $0 \neq v \in \text{aff}(N(\bar{z})) \cap N(\bar{z})^\perp$, giving a contradiction. The remaining equalities follow from Part (i) of the Theorem and

$$\text{lin}(T(\bar{z})) = N(\bar{z})^\perp = \text{aff}(N(\bar{z}))^\perp = T_S(\bar{z}).$$

(iv) From (iii), we have that the affine hulls of $\text{co}\{\nabla g_i(\bar{z}), i = 1, \dots, r\}$ and $N(\bar{z})$ are identical. The result follows from the definition of $\text{ri}(\cdot)$ and Definition 2(iii).

(v) Let $v \in T_S(\bar{z})$, and suppose for contradiction that $v^T \nabla^2 g_i(\bar{z}) v < 0$. There are sequences $v_j \rightarrow v$ and $\{t_j\}$ with $0 < t_j \in \mathbb{R}$, $t_j \rightarrow 0$ such that $\bar{z} + t_j v_j \in S \subset \Omega$, so $g_i(\bar{z} + t_j v_j) = 0$, $i = 1, \dots, r$. Since $\nabla g_i(\bar{z}) \in N(\bar{z})$, we have $\nabla g_i(\bar{z})^T t_j v_j \leq 0$, and so

$$0 = g_i(\bar{z} + t_j v_j) = g_i(\bar{z}) + \nabla g_i(\bar{z})^T (t_j v_j) + \frac{1}{2} t_j^2 v_j^T \nabla^2 g_i(\bar{z}) v_j,$$

where $\hat{v}_j \in [\bar{z}, \bar{z} + t_j v_j]$. Hence,

$$v_j^T \nabla^2 g_i(\hat{v}_j) v_j = -\frac{2}{t_j} v_j^T \nabla g_i(\bar{z}) \geq 0.$$

For j sufficiently large,

$$0 > \frac{1}{2} v^T \nabla^2 g_i(\bar{z}) v \geq v_j^T \nabla^2 g_i(\hat{v}_j) v_j \geq 0,$$

giving a contradiction. \square

The next result, which will be useful when we come to prove finite identification properties for constrained optimization algorithms, shows that the direct sum of an identifiable surface S and the relative interior of the normal cones along S , is a set that is open in \mathbb{R}^n . This property is analogous to that described for open facets in Theorem 2.8 of Burke and Moré [1].

LEMMA 2.2. *Suppose that S is as in Lemma 2.1 with $p \geq 2$. Define the set*

$$K = \{x + w \mid x \in S, w \in \text{ri}(N(x))\}.$$

For each $y \in K$, there is a $\delta \in (0, \epsilon(y))$ such that $y + \delta B \subset K$, that is, K is open in \mathbb{R}^n .

Proof. We start by finding $\delta > 0$ such that $u \in y + \delta B \Rightarrow P(u) \in S$. Let $\bar{y} = P(y)$, $r = r(y)$, $\epsilon = \epsilon(y)$ and g_i , $i = 1, \dots, r$ be chosen as in Definition 2. Let $\delta_1 \in (0, \epsilon)$ have the property that $(y + \delta_1 B) \cap \Omega = \emptyset$. By Definition 2(ii),(iv),(v), we know that there is $\lambda \in \mathbb{R}^r$ with $\lambda > 0$ such that

$$(5) \quad \begin{aligned} y - \bar{y} - \nabla g(\bar{y})\lambda &= 0, \\ g(\bar{y}) &= 0. \end{aligned}$$

From Definition 2(i) and Lemma 2.1(iii), we know that the matrix

$$\begin{bmatrix} I + \sum_{i=1}^r \lambda_i \nabla^2 g_i(\bar{y}) & \nabla g(\bar{y}) \\ \nabla g(\bar{y})^T & 0 \end{bmatrix}$$

is nonsingular and continuous with respect to \bar{y} and λ . We can now view \bar{y} and λ as functions of y in (5), and apply the implicit function theorem to obtain the following result: There is $\delta \in (0, \delta_1]$ such that, if $\|u - y\| \leq \delta$, the solution (\bar{u}, λ^u) of the system

$$(6) \quad \begin{aligned} u - \bar{u} - \nabla g(\bar{u})\lambda^u &= 0, \\ g_i(\bar{u}) &= 0 \end{aligned}$$

satisfies

$$(7) \quad \lambda^u > 0, \quad \|\bar{u} - \bar{y}\| < \epsilon.$$

Now (6) and (7) imply that (\bar{u}, λ^u) solves the problem

$$\min_{\bar{u}} \frac{1}{2} \|u - \bar{u}\|^2, \quad g(\bar{u}) = 0, \quad \|\bar{u} - \bar{y}\| < \epsilon$$

and so, by Definition 2(v), \bar{u} is the projection of u onto $S(\bar{y}; \epsilon)$. From (6), (7), and Definition 2(iii), we have that

$$u - \bar{u} \in \text{co} \{ \nabla g_i(\bar{u}), i = 1, \dots, r \} \subset N(\bar{u}).$$

Now $\bar{u} \in \Omega$, $u - \bar{u} \in N(\bar{u})$ and uniqueness of the projection onto a convex set imply that $\bar{u} = P(u)$.

Finally, since $\lambda^u > 0$, we have

$$u - \bar{u} \in \text{ri}(\text{co} \{ \nabla g_i(\bar{u}), i = 1, \dots, r \}).$$

Hence, by Lemma 2.1(iv),

$$u - \bar{u} \in \text{ri}(N(\bar{u})),$$

and so $u \in K$, as required. \square

In the following two results, we show how open facets and active index sets relate to identifiable surfaces.

THEOREM 2.3. *Let S be an open facet in Ω . Then S is a class- C^∞ identifiable surface.*

Proof. The case $S = \text{int}(\Omega)$ is trivially true. Consider $S \subset \partial\Omega$. Burke and Moré [1] show that any open facet S is the relative interior of a quasipolyhedral face. Hence, $N(\bar{y})$ and $T(\bar{y})$ are the same for all $\bar{y} \in S$, and

$$(8) \quad \text{aff}(S) = \bar{y} + \text{lin}(T(\bar{y}))$$

for all $\bar{y} \in S$.

Suppose, as in Definition 2, that we are given some y such that $\bar{y} = P(y) \in S$ and $y - \bar{y} \in \text{ri}(N(\bar{y}))$. Then there is a constant $\gamma > 0$ such that $(y - \bar{y}) + \gamma v \in \text{ri}(N(\bar{y}))$ for all $v \in \text{aff}(N(\bar{y}))$ with $\|v\| = 1$. Supposing that $\text{aff}(N(\bar{y}))$ has dimension r , we can choose unit vectors v_1, \dots, v_{r-1} , such that $\{v_1, \dots, v_{r-1}, y - \bar{y}\}$ is linearly independent in $\text{aff}(N(\bar{y}))$, and hence a spanning set. Now set

$$v_r = -\frac{1}{r-1}(v_1 + \dots + v_{r-1})$$

and

$$\hat{v}_i = y - \bar{y} + \gamma v_i, \quad i = 1, \dots, r.$$

Clearly, $\hat{v}_i \in \text{aff}(N(\bar{y}))$ and $\|\hat{v}_i - (y - \bar{y})\| \leq \gamma\|v_i\| \leq \gamma$, so $\hat{v}_i \in \text{ri}(N(\bar{y}))$, $i = 1, \dots, r$. Moreover, we can show that $\{\hat{v}_1, \dots, \hat{v}_r\}$ is linearly independent by the following argument: Suppose there are real coefficients μ_1, \dots, μ_r such that $\sum \mu_i \hat{v}_i = 0$. Then

$$\begin{aligned} 0 &= \sum_{i=1}^r \mu_i \hat{v}_i = \left(\sum_{i=1}^r \mu_i \right) (y - \bar{y}) + \gamma \sum_{i=1}^r \mu_i v_i \\ &= \left(\sum_{i=1}^r \mu_i \right) (y - \bar{y}) + \gamma \sum_{i=1}^{r-1} [\mu_i - \mu_r / (r-1)] v_i. \end{aligned}$$

By the original choice of v_1, \dots, v_{r-1} , we must have

$$\sum_{i=1}^r \mu_i = 0, \quad \mu_i = \mu_r / (r-1), \quad i = 1, \dots, r-1,$$

and it follows that $\mu_1 = \dots = \mu_r = 0$, as desired. Now define $g_i(z) = (z - \bar{y})^T \hat{v}_i$, $i = 1, \dots, r$. Conditions (i) and (ii) of Definition 2 are readily verified. Condition (iii) follows since $N(\bar{z})$ is constant for $\bar{z} \in S$, and $\hat{v}_i \in \text{ri}(N(\bar{z}))$, $i = 1, \dots, r$. Condition (iv) is verified by noting that

$$y - \bar{y} = \sum_{i=1}^{r-1} \frac{\hat{v}_i}{2(r-1)} + \frac{\hat{v}_r}{2} \in \text{co}\{\hat{v}_i, i = 1, \dots, r\}.$$

To prove condition (v), we first take $V = \text{aff}(S)$ in Definition 1 and note that, if $\bar{y} \in S$, there is $\epsilon > 0$ such that $\bar{z} \in \text{aff}(S) \cap (\bar{y} + \epsilon B) \Rightarrow \bar{z} \in S$. That is,

$$S(\bar{y}; \epsilon) = \{\bar{z} \mid \|\bar{z} - \bar{y}\| \leq \epsilon, \bar{z} \in \text{aff}(S)\} = \text{aff}(S) \cap (\bar{y} + \epsilon B).$$

However, by (8),

$$\text{aff}(S) = \bar{y} + \text{lin}(T(\bar{y})) = \bar{y} + N(\bar{y})^\perp,$$

and so,

$$\bar{z} \in \text{aff}(S) \Leftrightarrow (\bar{z} - \bar{y})^T \hat{v}_i = 0, \quad i = 1, \dots, r.$$

Hence,

$$S(\bar{y}, \epsilon) = \{\bar{z} \mid \|\bar{z} - \bar{y}\| \leq \epsilon, \quad g_i(\bar{z}) = 0, \quad i = 1, \dots, r\},$$

as required. \square

THEOREM 2.4. *Suppose that Ω is defined by (3) and (4), where g_i , $i = 1, \dots, m$ are C^1 functions. Suppose that for some set $\mathcal{A} \subset \{1, \dots, m\}$, the surface S defined by*

$$S = \{z \mid g_i(z) = 0, \quad i \in \mathcal{A}, \quad g_i(z) > 0, \quad i \notin \mathcal{A}\}$$

is a connected subset of Ω . Then S is a class- C^1 identifiable surface. Moreover, if $g_i \in C^p$ for $i \in \mathcal{A}$ and $p \geq 2$, then S is a class- C^p identifiable surface.

Proof. This follows trivially, by identifying g_i , $i \in \mathcal{A}$ with g_i , $i = 1, \dots, r$, in Definition 2. \square

We now consider smoothness of the projection operator $P(\cdot)$. The motivation for this comes from the work of Holmes [10] and Fitzpatrick and Phelps [6], who consider closed convex sets with smooth boundaries. In these papers, smoothness of the boundary is defined in terms of smoothness of the gauge function

$$\rho_\Omega(x) = \inf\{t > 0 \mid x \in t(\Omega - x_0) + x_0\}, \quad \text{for some } x_0 \in \text{int}(\Omega),$$

and the boundary of Ω is said to be C^p if ρ_Ω is C^p in some neighborhood of $\partial\Omega$. By showing that this definition is equivalent to a local C^p parametrization of the boundary, Holmes [10] essentially shows that a C^p boundary (by the definition above) is the same as a class- $C^{p,0}$ boundary, as defined in [9]. Hence, as discussed earlier, $\partial\Omega$ is a class- C^p identifiable surface.

Holmes proves the following result.

THEOREM 2.5 ([10, Thm. 2]). *If Ω has a C^p boundary, for $p \geq 2$, then the projection operator $P(\cdot)$ is C^{p-1} in $\mathbb{R}^n \setminus \Omega$, and $P'(y)$ is invertible in $\text{lin}(T(P(y)))$.*

Fitzpatrick and Phelps [6, Thm. 3.10] prove the converse.

The case of $p = 2$ is the most interesting. It is a classical result [5, p. 216] that, since P is Lipschitz continuous, it is differentiable almost everywhere. Below, we extend Theorem 2.5 to sets with *piecewise* smooth boundaries, by showing that class- C^p identifiable surfaces generate open regions in $\mathbb{R}^n \setminus \Omega$ in which P is C^{p-1} .

THEOREM 2.6. *Let S and K be as defined in Lemma 2.2, with $p \geq 2$. Then $P(\cdot)$ is C^{p-1} on K . Also, $P'(y)$ is invertible in $\text{lin}(T(P(y)))$.*

Proof. For any $y \in K$, we can choose $\epsilon > 0$ and $\delta > 0$ as in Lemma 2.2 such that, when $u \in y + \epsilon B$, $P(u)$ is also the projection of u onto the set $\{\bar{z} \mid g_i(\bar{z}) = 0, \quad i = 1, \dots, r, \quad \|\bar{z} - \bar{y}\| \leq \epsilon\}$. Hence, we can differentiate the system (5) with respect to y to obtain

$$(9) \quad \begin{bmatrix} I + \sum_{i=1}^r \lambda_i \nabla^2 g_i(\bar{y}) & \nabla g(\bar{y}) \\ \nabla g(\bar{y})^T & 0 \end{bmatrix} \begin{bmatrix} \frac{d\bar{y}}{dy} \\ \frac{d\lambda}{dy} \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix},$$

where $P'(y) = d\bar{y}/dy$. The first result follows immediately from (5) and the implicit function theorem (see, for example, Lang [11, p. 125]) by noting that the coefficient matrix in (9) is nonsingular.

For the second result, let $Z \in \mathbb{R}^{n \times (n-r)}$ be a matrix of full rank such that $\nabla g(\bar{y})^T Z = 0$. By Lemma 2.1(iii), the columns of Z span $\text{lin}(T(P(y)))$. The second equation in (9) implies that

$$P'(y) = \frac{d\bar{y}}{dy} = ZW^T$$

for some $W \in \mathbb{R}^{n \times (n-r)}$. Multiplying the first equation in (9) by Z^T , we find that

$$Z^T \left(I + \sum_{i=1}^r \lambda_i \nabla^2 g_i(\bar{y}) \right) ZW^T = Z^T$$

and so

$$(10) \quad P'(y) = Z \left[Z^T \left(I + \sum_{i=1}^r \lambda_i \nabla^2 g_i(\bar{y}) \right) Z \right]^{-1} Z^T.$$

It follows from (10) and Lemma 2.1(v) that $P'(y)$ has nonsingular projection onto $\text{lin}(T(P(y)))$. \square

We conjecture that the converse of this theorem is also true; that is, if there is an open connected region $K \subset \mathbb{R}^n \setminus \Omega$ such that $P(\cdot)$ is C^{p-1} on K , and $P'(y)$ is invertible in $\text{lin}(T(P(y)))$ for each $y \in K$, then $P(K)$ is a class- C^p identifiable surface. The continuity condition alone is not sufficient, as an example from Fitzpatrick and Phelps [6, p. 496] illustrates. Define

$$\Omega_4 = \{(\xi_1, \xi_2) \mid \xi_2 \geq |\xi_1| + \xi_1^{4/3}\}.$$

There is a corner in Ω_4 at $(0, 0)$, and the set has four maximal class- C^∞ identifiable surfaces: the corner, the interior, and the two edges. Tedious calculation shows that P is C^1 on $\mathbb{R}^n \setminus \Omega$, although $\partial\Omega$ is obviously not a class- C^2 surface. It can be shown that $P'(y) = 0$ along the lines $\{(\xi_1, \xi_2) \mid \xi_2 = -\xi_1, \xi_1 > 0\}$ and $\{(\xi_1, \xi_2) \mid \xi_2 = \xi_1, \xi_1 < 0\}$, and so the invertibility condition is not satisfied.

It is clear from (10) that the invertibility condition is related to the boundedness of the quantities $\lambda_i \nabla^2 g_i(\bar{y})$ on $\text{lin}(T(\bar{y}))$. Note that these quantities are invariant under scaling of the g_i s, that is, if g_i is replaced by αg_i , then λ_i becomes λ_i/α .

3. Finite identification in constrained optimization algorithms. We turn now to algorithms for solving the optimization problem (1).

In analyzing the gradient projection algorithm, we use the work of Dunn [4, §2], who provided a framework for proving capture results. Dunn states this algorithm as follows: Choosing constants γ_1 and γ_2 with $0 < \gamma_1 < \gamma_2 < 1$, and an initial iterate $x_0 \in \Omega$, set

$$(11) \quad x_{k+1} = P(x_k - \sigma_k \nabla F(x_k)),$$

where σ_k is chosen to satisfy

$$(12) \quad F(x_k) - F(P(x_k - \nabla F(x_k))) \geq \gamma_1 \Rightarrow \sigma_k = 1,$$

$$(13) \quad F(x_k) - F(P(x_k - \nabla F(x_k))) < \gamma_1 \Rightarrow \sigma_k \in (0, 1)$$

and

$$(14) \quad \gamma_1 \leq \frac{F(x_k) - F(P(x_k - \sigma_k \nabla F(x_k)))}{\nabla F(x_k)^T [x_k - P(x_k - \sigma_k \nabla F(x_k))]} \leq \gamma_2.$$

Gawande and Dunn [8] adapted Dunn's earlier work to prove capture and convergence results for *scaled* gradient projection algorithms and algebraic parametrizations (3) of the feasible set.

We start with a simple result that exploits openness of the set K of Lemma 2.2:

THEOREM 3.1. *Suppose that*

- (i) *Assumption 1 and (2) hold at some point x^* ;*
- (ii) *∇F is continuous at x^* ;*
- (iii) *$x^* \in S$, where S is a class- C^p identifiable surface of Ω with $p \geq 1$;*
- (iv) *there is $\bar{\sigma} > 0$ such that $\sigma_k \in [\bar{\sigma}, 1]$ for all k ; and*
- (v) *the sequence $\{x_k\}$ generated by (11)–(14) converges to x^* .*

Then $x_k \in S$ for all k sufficiently large.

Proof. Define the set K as in Lemma 2.2. Setting $y = x^* - \nabla F(x^*)$, we can apply Definition 2 to find $\delta > 0$ such that

$$x^* - \nabla F(x^*) + \delta B \subset K.$$

By construction of K , this implies that

$$x^* - \sigma \nabla F(x^*) + \sigma \delta B \subset K \quad \text{for all } \sigma \in [\bar{\sigma}, 1].$$

Now, choose \bar{k} such that, for all $k \geq \bar{k}$,

$$\|x_k - x^*\| + \|\nabla F(x_k) - \nabla F(x^*)\| \leq \bar{\sigma} \delta.$$

Then

$$\|[x_k - \sigma_k \nabla F(x_k)] - [x^* - \sigma_k \nabla F(x^*)]\| \leq \bar{\sigma} \delta,$$

and so

$$x_k - \sigma_k \nabla F(x_k) \in x^* - \sigma_k \nabla F(x^*) + \bar{\sigma} \delta B \subset x^* - \sigma_k \nabla F(x^*) + \sigma_k \delta B \subset K.$$

Hence, $x_{k+1} \in P(K) = S$, for $k \geq \bar{k}$. \square

Before proving the next result, we state second-order conditions and define some terms:

Assumption 2. Suppose that Ω satisfies Assumption 1 and that there is x^* that satisfies (2), such that $x^* \in S$, where S is a class- C^p identifiable surface of Ω with $p \geq 2$. Suppose that F is twice continuously differentiable in a neighborhood of x^* , and let g_i , $i = 1, \dots, r$ be as defined in Definition 2, for $y = x^* - \nabla F(x^*)$. Choose $\lambda^* \in \mathbb{R}^r$ such that $\lambda^* > 0$ and

$$(15) \quad -\nabla F(x^*) = \nabla g(x^*) \lambda^*,$$

and suppose that for all $h \in T_S(x^*)$,

$$h^T \left[\nabla^2 F(x^*) + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^*) \right] h \geq \alpha \|h\|^2, \quad \text{for some } \alpha > 0.$$

DEFINITION 3.

(i) x^* is a proper local minimizer of F in Ω if there is $\rho_1 > 0$ such that

$$x \in \Omega, 0 < \|x - x^*\| \leq \rho_1 \Rightarrow F(x) > F(x^*).$$

(ii) x^* is a stable fixed point for (11)–(14) if $-\nabla F(x^*) \in N(x^*)$, and there are $d_1 > 0, d_2 > 0$ such that

$$\|x_0 - x^*\| \leq d_1 \Rightarrow \|x_k - x^*\| \leq d_2 \quad \text{for all } k \geq 0.$$

(iii) x^* is a stable local attractor for (11)–(14) if it is a stable fixed point, and $d_1 > 0$ can be chosen so that

$$\|x_0 - x^*\| \leq d_1 \Rightarrow \lim_{k \rightarrow \infty} x_k = x^*.$$

The following theorem contains capture and convergence results like those proved in Gawande and Dunn [8, §4]. Here, we prove these results for the gradient projection method on the identifiable surface containing x^* ; in [8], the focus was on scaled gradient projection methods and active index sets for Ω defined by (3).

THEOREM 3.2. *Suppose that Assumption 2 holds. Then*

(i) *there are positive scalars ρ_1 and α_1 such that*

$$x \in \Omega, \|x - x^*\| \leq \rho_1 \Rightarrow F(x) - F(x^*) \geq \alpha_1 \|x - x^*\|^2;$$

(ii) *there are positive scalars ρ_2 and α_2 such that*

$$x \in S, \|x - x^*\| \leq \rho_2 \Rightarrow \|x - P(x - \nabla F(x))\| \geq \alpha_2 \|x - x^*\|,$$

that is, the defect $E(x) = x - P(x - \nabla F(x))$, restricted to S , has an isolated zero at x^ ;*

(iii) *given any $\bar{\sigma} > 0$, there is $\rho_3 = \rho_3(\bar{\sigma}) > 0$ such that*

$$\|x - x^*\| \leq \rho_3, \sigma \in [\bar{\sigma}, 1] \Rightarrow P(x - \sigma \nabla F(x)) \in S;$$

(iv) *x^* is a stable local attractor for the gradient projection algorithm, and the sequences $\{x_k\}$ that approach x^* eventually enter and remain in S .*

Proof. Throughout the proof, let ϵ denote $\epsilon(x^* - \nabla F(x^*))$.

(i) We show first that if $w = \nabla g(x^*)\mu$ with $\nabla g_i(x^*)^T w \leq c_1$ for $c_1 \geq 0$ and $i = 1, \dots, r$, then there is some $\tau_1 > 0$ such that

$$(16) \quad \nabla F(x^*)^T w \geq \tau_1 \|w\| + O(c_1).$$

Using the definition of λ^* from Assumption 2, we have

$$\begin{aligned} \nabla F(x^*)^T w &= -\lambda^{*T} \nabla g(x^*)^T \nabla g(x^*) \mu \\ &\geq (\min_i \lambda_i^*) \|\nabla g(x^*)^T \nabla g(x^*) \mu\| + O(c_1). \end{aligned}$$

Since $\nabla g(x^*)$ has full rank, and $\|\mu\| \geq \|w\|/\|\nabla g(x^*)\|$, there is τ_2 such that

$$\|\nabla g(x^*)^T \nabla g(x^*) \mu\| \geq \tau_2 \|\mu\| \geq \frac{\tau_2}{\|\nabla g(x^*)\|} \|w\|.$$

By setting

$$\tau_1 = (\min_i \lambda_i^*) \frac{\tau_2}{\|\nabla g(x^*)\|},$$

we obtain (16).

Now, given x in the vicinity of x^* , we seek vectors $v \in \mathbb{R}^n$ and $\mu \in \mathbb{R}^r$ such that

$$\begin{aligned} v + \nabla g(x^*)\mu &= x - x^* \\ g(x^* + v) &= 0. \end{aligned} \quad (17)$$

We can again apply the implicit function theorem to (17) to find $\bar{\rho}_1 > 0$ such that a solution v, μ exists for $\|x - x^*\| \leq \bar{\rho}_1$. Moreover, $\bar{\rho}_1$ can be chosen small enough that $\|v\| \leq \epsilon$, and hence, $x^* + v \in S \subset \Omega$. It follows that, since $\nabla g_i(x^* + v) \in N(x^* + v)$ for $i = 1, \dots, r$, and since $x \in \Omega$,

$$\nabla g_i(x^* + v)^T [x - (x^* + v)] \leq 0, \quad i = 1, \dots, r.$$

Writing $w = \nabla g(x^*)\mu = x - (x^* + v)$, we have

$$\nabla g_i(x^*)^T w = \nabla g_i(x^* + v)^T w + O(\|v\|\|w\|) \leq O(\|v\|\|w\|), \quad i = 1, \dots, r.$$

Application of (16) shows that

$$(18) \quad \nabla F(x^*)^T w \geq \tau_1 \|w\| + O(\|v\|\|w\|).$$

Now consider the v component. Since $x^* + v \in S$, and since (15) holds,

$$\begin{aligned} F(x^* + v) - F(x^*) &= [F(x^* + v) + \lambda^{*T} g(x^* + v)] - [F(x^*) + \lambda^{*T} g(x^*)] \\ (19) \quad &= \frac{1}{2} v^T \left[\nabla^2 F(x^* + \beta_1 v) + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^* + \beta_1 v) \right] v \end{aligned}$$

for some $\beta_1 \in (0, 1)$. Since $\nabla g_i(x^*)^T v = O(\|v\|^2)$, we can choose $\bar{\rho}_2 \in (0, \bar{\rho}_1]$ such that when $\|x - x^*\| \leq \bar{\rho}_2$, v is close enough to $T_S(x^*)$ and $\|v\|$ is small enough that

$$v^T \left[\nabla^2 F(x^* + \beta_1 v) + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^* + \beta_1 v) \right] v \geq \frac{\alpha}{2} \|v\|^2,$$

for all $\beta_1 \in [0, 1]$. Hence, from (19),

$$(20) \quad F(x^* + v) - F(x^*) \geq \frac{\alpha}{4} \|v\|^2.$$

By using (18) and (20), we can now write that, for $x \in \Omega \cap (x^* + \bar{\rho}_2 B)$,

$$\begin{aligned} F(x) - F(x^*) &= F(x^* + v + w) - F(x^* + v) + F(x^* + v) - F(x^*) \\ &= \nabla F(x^*)^T w + O(\|v\|\|w\| + \|w\|^2) + F(x^* + v) - F(x^*) \\ &\geq \tau_1 \|w\| + \frac{\alpha}{4} \|v\|^2 + O(\|v\|\|w\| + \|w\|^2) \\ (21) \quad &\geq \tau_1 \|w\| + \frac{\alpha}{4} \|v\|^2 - c_2 (\|v\|\|w\| + \|w\|^2), \end{aligned}$$

where $c_2 > 0$ is some constant. Now, choose a constant $\delta_1 > 0$ such that

$$(22) \quad c_2 (\delta_1^2 + \delta_1) \leq \frac{\alpha}{8},$$

and define $\rho_1 \in (0, \bar{\rho}_2]$ such that both of the following conditions are satisfied:

$$(23) \quad x \in \Omega \cap (x^* + \rho_1 B) \Rightarrow \|w\| \leq \frac{\tau_1 \delta_1}{2c_2(1 + \delta_1)},$$

$$(24) \quad \rho_1 \leq \frac{4\tau_1 \delta_1 (1 + \delta_1)}{\alpha}.$$

In the case $\|w\| \geq \delta_1 \|v\|$, we have

$$(25) \quad \|x - x^*\| \leq \|w\| + \|v\| \leq \left(1 + \frac{1}{\delta_1}\right) \|w\|.$$

Also, from (21),

$$(26) \quad F(x) - F(x^*) \geq \tau_1 \|w\| - c_2 \left(1 + \frac{1}{\delta_1}\right) \|w\|^2.$$

Now, from (23), (25), and (26), we have that

$$F(x) - F(x^*) \geq \frac{\tau_1}{2} \|w\| \geq \frac{\tau_1 \delta_1}{2(1 + \delta_1)} \|x - x^*\| \geq \frac{\tau_1 \delta_1}{2(1 + \delta_1)\rho_1} \|x - x^*\|^2,$$

for $x \in \Omega \cap (x^* + \rho_1 B)$. Application of (24) yields that

$$(27) \quad F(x) - F(x^*) \geq \frac{\alpha}{8(1 + \delta_1)^2} \|x - x^*\|^2.$$

In the remaining case $\|w\| < \delta_1 \|v\|$, we find from (21) and (22) that

$$F(x) - F(x^*) \geq \frac{\alpha}{4} \|v\|^2 - c_2(\delta_1^2 + \delta_1) \|v\|^2 \geq \frac{\alpha}{8} \|v\|^2.$$

Also,

$$\|x - x^*\| \leq \|v\| + \|w\| < (1 + \delta_1) \|v\|,$$

and hence, (27) still applies. The result follows by setting

$$\alpha_1 = \frac{\alpha}{8(1 + \delta_1)^2}.$$

(ii) By setting $y = x^* - \nabla F(x^*)$ in Lemma 2.2, we can choose $\delta \in (0, \epsilon]$ such that $P(x^* - \nabla F(x^*) + \delta B) \subset S$. Now, there is a $\bar{\rho}_1 \in (0, \epsilon]$ such that

$$\|x - x^*\| \leq \bar{\rho}_1 \Rightarrow \|[x - \nabla F(x)] - [x^* - \nabla F(x^*)]\| \leq \delta,$$

and hence, $\hat{x} = P(x - \nabla F(x)) \in S$. By contractivity of $P(\cdot)$, $\|\hat{x} - x^*\| = \|P(x - \nabla F(x)) - P(x^* - \nabla F(x^*))\| \leq \delta \leq \epsilon$. It therefore follows from Definition 2 (v) that \hat{x} solves the projection subproblem

$$(28) \quad \min_{\hat{x}} \frac{1}{2} \|\hat{x} - (x - \nabla F(x))\|_2^2, \quad g(\hat{x}) = 0.$$

When $x = x^*$, then $\hat{x} = x^*$, and (15) holds. By the implicit function theorem, $\bar{\rho}_2 \in (0, \bar{\rho}_1]$ can be chosen small enough that there is λ such that, in fact,

$$(29) \quad \|x - x^*\| \leq \bar{\rho}_2 \Rightarrow [x - \nabla F(x)] - \hat{x} = \nabla g(\hat{x})\lambda,$$

with

$$\|\hat{x} - x^*\| = O(\|x - x^*\|), \quad \|\lambda - \lambda^*\| = O(\|x - x^*\|), \quad \lambda > 0.$$

Since $\|x - x^*\| \leq \bar{\rho}_1 \leq \epsilon$, we also have by Definition 2(v) that $g_i(x) = 0$, $i = 1, \dots, r$.

Let $Z \in \mathbb{R}^{n \times (n-r)}$ be an orthonormal matrix whose columns span the subspace $T_S(x^*)$. By using a Taylor series expansion of g about x^* , it is easy to show that there are vectors $\eta, \hat{\eta} \in \mathbb{R}^{n-r}$ and $\zeta, \hat{\zeta} \in \mathbb{R}^r$ such that

$$(30) \quad x - x^* = Z\eta + \nabla g(x^*)\zeta,$$

$$(31) \quad x - \hat{x} = Z\hat{\eta} + \nabla g(x^*)\hat{\zeta},$$

where $\|\zeta\| = O(\|x - x^*\|^2)$ and $\|\hat{\zeta}\| = O(\|x - x^*\|^2) + O(\|\hat{x} - x^*\|^2) = O(\|x - x^*\|^2)$. From the second-order conditions and boundedness of $\nabla^2 g_i$ in a neighborhood of x^* there exists a constant $c_2 > 0$ such that

$$(32) \quad \eta^T Z^T \left[\nabla^2 F(x^*) + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^*) \right] Z\eta \geq \alpha \|\eta\|^2,$$

and

$$(33) \quad \left\| Z^T \left[I + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^*) \right] Z\eta \right\| \leq c_2 \|\eta\|,$$

for all $\eta \in \mathbb{R}^{n-r}$. It follows trivially from (32) that

$$(34) \quad \|Z^T \left[\nabla^2 F(x^*) + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^*) \right] Z\eta\| \geq \alpha \|\eta\|.$$

Now, from (29),

$$\begin{aligned} x - \hat{x} &= \nabla F(x) + \nabla g(\hat{x})\lambda \\ \Rightarrow x - \hat{x} &= \nabla F(x) + \nabla g(x)\lambda^* + [\nabla g(\hat{x}) - \nabla g(x)]\lambda + \nabla g(x)[\lambda - \lambda^*] \\ \Rightarrow x - \hat{x} &= \left[\nabla^2 F(x^{(1)}) + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^{(1)}) \right] (x - x^*) + \\ &\quad \sum_{i=1}^r \lambda_i \nabla^2 g_i(x^{(2)})(\hat{x} - x) + \nabla g(x^*)[\lambda - \lambda^*] + O(\|x - x^*\|^2), \end{aligned}$$

for $x^{(1)} \in [x, x^*]$ and $x^{(2)} \in [\hat{x}, x]$. Premultiplying this equation by Z^T , and using (30) and (31), we obtain

$$\begin{aligned} & Z^T \left[I + \sum_{i=1}^r \lambda_i \nabla^2 g_i(x^{(2)}) \right] [Z\hat{\eta} + \nabla g(x^*)\hat{\zeta}] \\ &= Z^T \left[\nabla^2 F(x^{(1)}) + \sum_{i=1}^r \lambda_i^* \nabla^2 g_i(x^{(1)}) \right] [Z\eta + \nabla g(x^*)\zeta] + O(\|x - x^*\|^2). \end{aligned}$$

Now, from (33) and (34), we can choose $\bar{\rho}_3 \in (0, \bar{\rho}_2]$ small enough that, for $\|x - x^*\| \leq \bar{\rho}_3$,

$$\begin{aligned} 2c_2 \|\hat{\eta}\| &\geq \left\| Z^T \left[I + \sum_{i=1}^r \lambda_i \nabla^2 g_i(x^{(2)}) \right] Z \hat{\eta} \right\| \\ &\geq O(\|\hat{\zeta}\|) + O(\|\zeta\|) + O(\|x - x^*\|^2) + \frac{\alpha}{2} \|\eta\|. \end{aligned}$$

Since

$$\begin{aligned} \|\hat{\eta}\| &= \|x - \hat{x}\| + O(\|x - x^*\|^2), \\ \|\eta\| &= \|x - x^*\| + O(\|x - x^*\|^2), \end{aligned}$$

there is a constant $c_3 > 0$ such that

$$\begin{aligned} 2c_2 \|x - \hat{x}\| &\geq \frac{\alpha}{2} \|x - x^*\| - c_3 \|x - x^*\|^2 \\ \Rightarrow \|x - \hat{x}\| &\geq \frac{\alpha}{4c_2} \|x - x^*\| \left[1 - \frac{2c_3}{\alpha} \|x - x^*\| \right]. \end{aligned}$$

Now, choosing $\rho_2 = \min(\bar{\rho}_3, \alpha/(4c_3))$, the desired result follows, with $\alpha_2 = \alpha/(8c_2)$.

(iii) The proof of this part is identical to that of Theorem 3.1, and hence, is omitted.

(iv) This follows from Theorem 2.1 of Dunn [4], after we make the following observations. Part (i) of this theorem implies that x^* is a uniformly proper local minimizer of F in Ω . The fact that $F \in C^2$ in a neighborhood of x^* means that it is possible to choose a $\bar{\sigma} \in (0, 1)$ such that, for x_k in this neighborhood, any σ_k satisfying (12)–(14) lies in $[\bar{\sigma}, 1]$. \square

We turn now to Newton-like methods for (1). Here, an initial iterate $x_0 \in \Omega$ is chosen, and for each $k \geq 0$, the following subproblem is solved to find a search direction p_k :

$$(35) \quad \min_{p_k} \nabla F(x_k)^T p_k + \frac{1}{2} p_k^T B_k p_k, \quad x_k + p_k \in \Omega.$$

A steplength $\sigma_k \in [0, 1]$ is chosen, usually with the help of some “sufficient decrease” criterion, and the next iterate is obtained by setting

$$(36) \quad x_{k+1} = x_k + \sigma_k p_k.$$

A simple result, similar to Theorem 3.1, follows.

THEOREM 3.3. *Suppose that*

- (i) *Assumption 1 and (2) hold at some point x^* ;*
- (ii) *$\nabla F(x)$ is continuous at x^* ;*
- (iii) *$x^* \in S$, where S is some class- C^p identifiable surface of Ω with $p \geq 1$; and*
- (iv) *$x_k \rightarrow x^*$ and $p_k \rightarrow 0$ as $k \rightarrow \infty$, and $\{\|B_k\|\}$ is bounded.*

Then $x_k + p_k \in S$ for all k sufficiently large.

Proof. As in Lemma 2.2, we can find a set $K \subset \mathbb{R}^n \setminus \Omega$ with $P(K) \subset S$, and a scalar $\delta > 0$ such that

$$x^* - \nabla F(x^*) + \delta B \subset \text{int}(K).$$

First-order conditions for (35) are that

$$-\nabla F(x_k) - B_k p_k \in N(x_k + p_k) \Leftrightarrow x_k + p_k = P(x_k + p_k - \nabla F(x_k) - B_k p_k).$$

Now,

$$\begin{aligned} & \| (x_k + p_k - \nabla F(x_k) - B_k p_k) - (x^* - \nabla F(x^*)) \| \\ & \leq \| x_k - x^* \| + \| \nabla F(x_k) - \nabla F(x^*) \| + (1 + \| B_k \|) \| p_k \|. \end{aligned}$$

We can choose \bar{k} large enough that, for $k \geq \bar{k}$, the right-hand side of the above inequality does not exceed δ . Then $x_k + p_k \in S$, as required. \square

Finally, we prove a capture and convergence result for Newton's method, which makes use of the second-order conditions in Assumption 2.

THEOREM 3.4. *Suppose that Assumption 2 holds, and that, in addition, $\nabla^2 F(x)$ is Lipschitz continuous in a neighborhood of x^* . Let $B_k = \nabla^2 F(x_k)$ in (35). Then there are positive constants ρ_4 and α_4 such that, if $x_0 \in \Omega \cap (x^* + \rho_4 B)$ and $\sigma_k \equiv 1$ for all $k \geq 0$, then the algorithm (35), (36) generates a sequence $\{x_k\}$ such that*

$$\|x_{k+1} - x^*\| \leq \alpha_4 \|x_k - x^*\|^2 \quad \text{for all } k \geq 0.$$

In addition, $x_k \in S$ for all k sufficiently large.

Proof. The proof of the first part follows from results of Dunn [3, Thm. 3.1, Note 3.1] provided that we find positive constants $\bar{\alpha}_1$ and $\bar{\rho}_1$ such that

$$(37) \quad \nabla F(x^*)^T (x - x^*) + \frac{1}{2} (x - x^*)^T \nabla^2 F(x^*) (x - x^*) \geq \bar{\alpha}_1 \|x - x^*\|^2$$

for all $x \in \Omega \cup (x^* + \bar{\rho}_1 B)$.

Suppose we choose $\bar{\rho}_1$ small enough that F is twice Lipschitz continuously differentiable on the open ball $x^* + \bar{\rho}_1 B$, with Lipschitz constant L and, in addition, that

$$\bar{\rho}_1 \leq \min(\rho_1, \alpha_1/L),$$

where α_1 and ρ_1 are the constants from Theorem 3.2(i). For $\|x - x^*\| \leq \bar{\rho}_1$, we have

$$\begin{aligned} & \nabla F(x^*)^T (x - x^*) + \frac{1}{2} (x - x^*)^T \nabla^2 F(x^*) (x - x^*) \\ & \geq F(x) - F(x^*) - \frac{1}{2} L \|x - x^*\|^3 \\ & \geq [\alpha_1 - \frac{1}{2} L \|x - x^*\|] \|x - x^*\|^2 \\ & \geq \frac{1}{2} \alpha_1 \|x - x^*\|^2, \end{aligned}$$

and so (37) is satisfied if we set $\bar{\alpha}_1 = \frac{1}{2} \alpha_1$.

The final statement in the theorem follows from Theorem 3.3. \square

Acknowledgment. I am grateful to the referees of this paper for their perceptive comments.

REFERENCES

- [1] J. V. BURKE AND J. J. MORÉ, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.
- [2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [3] J. C. DUNN, *Newton's method and the Goldstein step-length rule for constrained minimization problems*, SIAM J. Control Optim., 18 (1980), pp. 659–674.

- [4] J. C. DUNN, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–216.
- [5] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.
- [6] S. FITZPATRICK AND R. R. PHELPS, *Differentiability of the metric projection in Hilbert space*, Trans. Amer. Math. Soc., 270 (1982), pp. 483–501.
- [7] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.
- [8] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Appl. Math. Optim., 17 (1988), pp. 103–119.
- [9] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
- [10] R. B. HOLMES, *Smoothness of certain metric projections on Hilbert space*, Trans. Amer. Math. Soc., 183 (1973), pp. 87–100.
- [11] S. LANG, *Analysis II*, Addison-Wesley, Reading, MA, 1969.

A DUAL APPROACH TO LINEAR INVERSE PROBLEMS WITH CONVEX CONSTRAINTS*

LEE C. POTTER[†] AND K. S. ARUN[‡]

Abstract. A simple constraint qualification is developed and used to derive an explicit solution to a constrained optimization problem in Hilbert space. A finite parameterization is obtained for the minimum norm element in the intersection of a linear variety of finite co-dimension and a closed convex constraint set. The result extends previous duality theorems for convex cone set constraints. A fixed point iteration is presented for computing the parameters and yields a least-squares solution when the variety and constraint set have empty intersection. Proofs rely on nearest-point projections onto convex sets and the properties of monotone, firmly nonexpansive, and averaged mappings.

Key words. constrained optimization, semi-infinite convex program, constraint qualification, successive approximations, nearest-point projection, monotone operator

AMS(MOS) subject classifications. 49A, 49B, 49D

1. Introduction. The recovery of a signal from linear measurements and prior information is a central problem in signal analysis and remote sensing applications ranging from tomographic imaging and radio astronomy to well logging and respiratory physiology. Simplicity and generality are sought in characterizing and computing signals that successfully reflect available prior knowledge. To this end, the signal is abstractly represented as an element of a Hilbert space, and each known property of the signal is incorporated by restricting the reconstructed signal to lie in a specified closed convex set. In addition, the requirement that the signal be consistent with a finite number of linear measurements defines a linear variety of finite co-dimension. The intersection of this variety and the convex constraint set is termed the *feasible set* of signals. In this paper, the recovery task is formulated as the infinite-dimensional programming problem of determining the feasible signal closest to a specified nominal signal.

The desired signal is shown to admit a dual parameterization by exploiting the properties of monotone operators and nearest-point mappings onto closed convex sets. The parameter vector is seen to be a fixed point of a nonlinear, monotone, firmly nonexpansive operator in a finite-dimensional space; these properties lead both to a novel constraint qualification assuring the existence of the parameters and to iterative computational schemes. Convergence to a least-squares fit of the linear measurements is obtained when the feasible set is empty. The duality result does not require the constraint sets to have interior and allows direct derivation of the optimal L_2 solution in [8]. In addition, more recent L_p optimization results [2], [9] that likewise eschew the traditional Slater-type constraint qualification are extended, in Hilbert space, from the special case of a convex cone to general convex set constraints.

2. Problem formulation. Let \mathcal{S} be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$. By the Riesz representation theorem, any N continuous, linear measurement functionals on \mathcal{S} may be expressed by inner products with *measurement signals*

* Received by the editors December 10, 1990; accepted for publication (in revised form) February 28, 1992. This work was supported in part by the National Science Foundation grant MIP-9111044 and in part by a grant from the Strategic Defense Initiative Organization/Innovative Science and Technology managed through the U.S. Army Research Office contract DAAL03-86-K0111.

[†] Department of Electrical Engineering, The Ohio State University, Columbus, Ohio 43210.

[‡] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109.

g_1, g_2, \dots, g_N in \mathcal{S} . Accordingly, define the mapping A from \mathcal{S} into Euclidean N -space \mathbb{R}^N by

$$Ax = [\langle x, g_1 \rangle, \dots, \langle x, g_N \rangle]^t,$$

where $[\cdot]^t$ denotes vector transpose. For a given $\beta \in \mathbb{R}^N$, the set of all x satisfying $Ax = \beta$ is a linear variety of co-dimension not exceeding N . The adjoint operator A^* maps a vector $\theta \in \mathbb{R}^N$ with k^{th} entry θ_k to the signal $A^*\theta = \sum_{k=1}^N \theta_k g_k$. Thus, $A^*\theta$ is very simply a linear combination of the N measurement signals, and the range of A^* is the finite-dimensional subspace $\mathcal{G} \subset \mathcal{S}$ spanned by the measurement signals: $\text{range}(A^*) = \mathcal{G} = \text{span}\{g_1, \dots, g_N\}$. Let Π be the orthogonal projection onto \mathcal{G} . The orthogonal complement of \mathcal{G} is the null space of A , denoted $\ker(A)$; the linear variety $\{x : Ax = \beta\}$ is a translate of $\ker(A)$ and is therefore a closed convex set.

Let $\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_M$ be closed convex sets with nonempty intersection \mathcal{K} . The set \mathcal{K} is referred to as the *constraint set*; \mathcal{K} may be infinite-dimensional and is not assumed to have interior. For a fixed measurement vector β the *feasible set* \mathcal{F} is defined to be the intersection of the variety $\{x : Ax = \beta\}$ with the constraint set \mathcal{K} . That is, \mathcal{F} is the closed and convex set $\{x \in \mathcal{K} : Ax = \beta\}$. Finally, let \mathcal{E} denote the *extendible set* in \mathbb{R}^N defined to consist of all measurement vectors β for which the associated \mathcal{F} is nonempty.

The recovery problem is to characterize and compute the signal in the feasible set \mathcal{F} closest to a specified nominal signal. Without loss of generality, the nominal signal, x_{nom} , is the origin: for $x_{nom} \neq 0$, the data vector is replaced by $\beta - Ax_{nom}$, and the constraint set is translated by $-x_{nom}$. This constrained inverse problem is concisely written

$$(P) \quad \min_{x \in \mathcal{K}} \|x\| \quad \text{subject to} \quad Ax = \beta.$$

The special case in which \mathcal{K} is a convex cone is considered in [2], [8], [9], and [25], and subspace or linear variety constraints are considered in [3], [14]. Problem (P) is the linear inverse problem $Ax = \beta$ with the additional convex set constraint $x \in \mathcal{K}$.

Were the distinction between the closed convex data constraint $Ax = \beta$ and the set constraint \mathcal{K} to be abandoned, the minimum norm element of the feasible set, \mathcal{F} , would be trivially characterized by the projection of the origin onto closed convex set \mathcal{F} . However, this conceptual approach is undesirable since the aim is to explicitly determine solutions. First, to combine the data constraint with the set \mathcal{K} forfeits the structural advantage afforded by the finite co-dimensionality of the linear variety. Second, the nearest-point projection operator onto \mathcal{F} may not be computable in a tractable manner; the set \mathcal{K} , on the other hand, typically arises from physically meaningful constraints that give rise to an easily implemented nearest-point projection operator onto \mathcal{K} . Third, the distinction between the data matching and set constraints allows for the computation of a least-squares solution when measurement noise renders the feasible set empty.

3. An optimality condition. In the absence of the constraints imposed by the convex set \mathcal{K} , the projection theorem, e.g., [13] simply and elegantly characterizes the minimum norm element of the variety $\{x : Ax = \beta\}$ as a linear combination $\sum_{k=1}^N \theta_k g_k$, where the parameters $\theta \in \mathbb{R}^N$ are determined by the normal equations. In a similar manner, the constraints embodied by \mathcal{K} are incorporated, and a particularly simple and geometrically appealing optimization result for (P) is obtained. The following theorem establishes a parsimonious parameterization of the solution \hat{x} .

First, two basic facts are reviewed for closed convex sets in a Hilbert space.

LEMMA 1. *Let \mathcal{K} denote any closed convex subset of a Hilbert space \mathcal{S} . Then there exists a unique $y \in \mathcal{K}$ such that $\inf_{z \in \mathcal{K}} \|x - z\| = \|x - y\|$.*

This correspondence is denoted by $y = P_{\mathcal{K}}(x)$, where $P_{\mathcal{K}} : \mathcal{S} \mapsto \mathcal{K}$ is said to be the *nearest-point projection operator*, or simply the *projection*, of \mathcal{S} onto the closed convex set \mathcal{K} . The operator $P_{\mathcal{K}}$ is linear if and only if \mathcal{K} is a subspace.

LEMMA 2. *Let \mathcal{K} be a closed convex subset of \mathcal{S} . Then the following are equivalent:*

- (a) $P_{\mathcal{K}}(x) = y$
- (b) $\|x - y\| \leq \|x - z\|$ for all $z \in \mathcal{K}$
- (c) $\langle x - y, z - y \rangle \leq 0$ for all $z \in \mathcal{K}$.

THEOREM 1. *If there exists $\theta \in \mathbb{R}^N$ such that $\beta = AP_{\mathcal{K}}A^*(\theta)$, then $\hat{x} = P_{\mathcal{K}}A^*(\theta)$ is the unique solution to (P).*

Proof [18], [20]. The feasible set $\mathcal{F} := \mathcal{K} \cap \{x : Ax = \beta\}$ is closed and convex, and the existence of a unique minimum norm element follows from Lemma 1, provided \mathcal{F} is nonempty. Let $y := A^*\theta$ where θ is the parameter vector of the hypothesis. It must be shown that

$$\inf_{x \in \mathcal{F}} \|x\| = \|P_{\mathcal{K}}(y)\|.$$

From Lemma 2, it suffices to show that $\langle P_{\mathcal{K}}(y), x - P_{\mathcal{K}}(y) \rangle \geq 0$ for all $x \in \mathcal{F}$. To this end, let x denote an arbitrary element of \mathcal{F} . Now write $P_{\mathcal{K}}(y)$ as $y - (y - P_{\mathcal{K}}(y))$ to yield

$$\langle P_{\mathcal{K}}(y), x - P_{\mathcal{K}}(y) \rangle = \langle y, x - P_{\mathcal{K}}(y) \rangle - \langle y - P_{\mathcal{K}}(y), x - P_{\mathcal{K}}(y) \rangle.$$

First, observe that $\langle y, x - P_{\mathcal{K}}(y) \rangle = 0$ since $y \in \mathcal{G} = \text{range}(A^*)$ and $Ax = \beta = AP_{\mathcal{K}}(y)$ implies $(x - P_{\mathcal{K}}(y)) \in \ker(A)$. Turning to the second term, observe from Lemma 2 that $\langle y - P_{\mathcal{K}}(y), x - P_{\mathcal{K}}(y) \rangle \leq 0$ for all $x \in \mathcal{K}$. In particular, \mathcal{F} is a subset of \mathcal{K} , so the inequality holds for all x in \mathcal{F} . Hence, $\langle P_{\mathcal{K}}(y), x - P_{\mathcal{K}}(y) \rangle \geq 0$ for all x in \mathcal{F} . \square

The result in Theorem 1 is a nonlinear generalization of the classical projection theorem, which follows as a simple corollary.

COROLLARY 1. *For $\mathcal{K} = \mathcal{S}$ and $\mathcal{F} \neq \emptyset$, the solution \hat{x} to (P) is given by $A^*\theta$, where θ satisfies the normal equations $AA^*\theta = \beta$.*

Figure 1 provides an illustration of Theorem 1 in the Euclidean plane and, although depicting the degenerate case of $\mathcal{S} = \mathbb{R}^2$, illuminates the similarities between Theorem 1 and the projection theorem. The minimum norm element x_{mn} of the variety $\{x : Ax = \beta\}$ is the orthogonal projection of the origin onto the variety. Thus, $x_{mn} = \sum_{k=1}^N \theta_k g_k$, where the coefficients θ_k are uniquely specified by the linear equations in Corollary 1. However, the minimum norm solution lies outside the constraint set \mathcal{K} , in general. Yet, the constrained minimum norm element \hat{x} is found in an analogous manner: \hat{x} is the nearest-point projection onto \mathcal{K} of an element $A^*\hat{\theta} = \sum_{k=1}^N \hat{\theta}_k g_k$ in \mathcal{G} , where the parameter vector $\hat{\theta}$ is determined by the equations in Theorem 1. Thus, in order to constrain the minimum norm solution to lie in the constraint set \mathcal{K} , the linear normal equations $AA^*\theta = \beta$ are replaced by the nonlinear equations $AP_{\mathcal{K}}A^*(\hat{\theta}) = \beta$, and the solution $x_{mn} = A^*\theta$ is replaced by $\hat{x} = P_{\mathcal{K}}A^*(\hat{\theta})$.

4. A constraint qualification. The hypothesis of Theorem 1 requires the existence of a solution $\hat{\theta}$ to a nonlinear system of equations, and the optimal signal \hat{x} is then parameterized by this solution via $\hat{x} = P_{\mathcal{K}}A^*(\hat{\theta})$. For a nonempty feasible set,

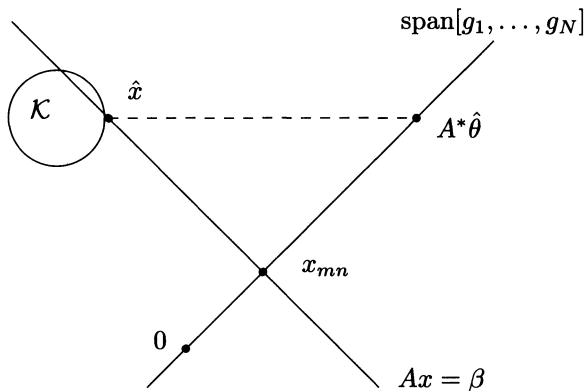


FIG. 1. The minimum norm feasible signal is the projection onto the constraint set of an element in the span of the measurement signals.

existence and uniqueness of a solution, \hat{x} , to (P) follow from Lemma 1. Therefore, the statement of the theorem immediately raises the questions: When does the representation of \hat{x} by $\hat{\theta}$ exist? Is the representation unique? How may it be computed? To address the issues of existence and uniqueness requires an investigation of the ranges of the nonlinear operator $AP_{\mathcal{K}}A^* : \mathbb{R}^N \mapsto \mathbb{R}^N$ and its set-valued inverse. That is, there exists a solution to $AP_{\mathcal{K}}A^*(\theta) = \beta$ if and only if β is in the range of $AP_{\mathcal{K}}A^*$, and the solution is unique if and only if $(AP_{\mathcal{K}}A^*)^{-1}(\beta)$ is single valued. Pertinent properties of these ranges are derived in this section by making use of their finite dimensionality and utilizing results from the theory of monotone operators. These properties are then used both to establish a novel constraint qualification (Cor. 2), which gives a condition on the data vector β to ensure the existence of a parameterization and to characterize uniqueness. The third issue, computation, is deferred to §5, where the solution of $AP_{\mathcal{K}}A^*(\theta) = \beta$ is viewed as a nonlinear fixed point problem.

A set \mathcal{M} in the Cartesian product $\mathbb{R}^N \times \mathbb{R}^N$ is said to be *monotone*, e.g., [27], if

$$\langle x^* - y^*, x - y \rangle \geq 0 \quad \forall (x, x^*), (y, y^*) \in \mathcal{M}.$$

A *maximal monotone* set is one not properly contained in another monotone set. A (possibly set valued) mapping $f : \mathbb{R}^N \mapsto 2^{\mathbb{R}^N}$ is called a *monotone operator* if its graph $\{(x, x^*) | x^* \in f(x)\}$ is a monotone set in $\mathbb{R}^N \times \mathbb{R}^N$; the operator is said to be *maximal monotone* if its graph is a maximal monotone set. The operator f^{-1} is defined as the mapping which has as its graph the set $\{(x^*, x) | (x, x^*) \in \text{graph of } f\}$. Since monotonicity is invariant under transposition of the domain and range of a map, f and f^{-1} are simultaneously monotone or maximal monotone. In the sequel, set-valued mappings will be viewed as multifunctions, and the notation $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ will be employed.

The properties of maximal monotone operators in finite-dimensional spaces and convex sets in \mathbb{R}^N are combined to guarantee that for any data vector β in the relative interior of the extendible set \mathcal{E} there exists a parameter vector $\hat{\theta}$ providing the representation $\hat{x} = P_{\mathcal{K}}A^*(\hat{\theta})$. The requisite properties are established as two brief lemmas; the resulting theorem gives the desired constraint qualification as an immediate corollary.

LEMMA 3. *The operators $AP_{\mathcal{K}}A^* : \mathbb{R}^N \mapsto \mathbb{R}^N$ and $\Pi P_{\mathcal{K}}\Pi : \mathcal{S} \mapsto \mathcal{S}$ are maximal monotone operators.*

Proof. From the linearity of A and the definition of the adjoint A^* , it follows that

$$\begin{aligned}\langle AP_{\mathcal{K}}A^*(x) - AP_{\mathcal{K}}A^*(y), x - y \rangle &= \langle P_{\mathcal{K}}A^*(x) - P_{\mathcal{K}}A^*(y), A^*(x - y) \rangle \\ &= \langle P_{\mathcal{K}}(x^*) - P_{\mathcal{K}}(y^*), x^* - y^* \rangle,\end{aligned}$$

where $x^* = A^*x$ and $y^* = A^*y$. Immediately, this inner product is nonnegative since the projection $P_{\mathcal{K}}$ is monotone [7]. Next, since $AP_{\mathcal{K}}A^*$ is continuous and defined for all x in \mathfrak{R}^N , it is maximal monotone [15]. The proof for $\Pi P_{\mathcal{K}}\Pi$ is identical. \square

LEMMA 4 ([17]). *The closure of the range of a maximal monotone operator is a convex set.*

As defined above, the set of all data vectors β that can be generated by measuring some signal x from the constraint set \mathcal{K} is termed the *extendible set*. This set, $\mathcal{E} := \{\beta \in \mathfrak{R}^N : \beta = Ax, x \in \mathcal{K}\}$, is convex (immediately from the linearity of A) but not necessarily closed. Lemmas 3 and 4 are used to establish that the extendible set and the range of $AP_{\mathcal{K}}A^*$ are the same to within closure.

THEOREM 2. *The closure of the extendible set is equal to the closure of the set of all measurement vectors obtainable from parameterized signals of the form $x = P_{\mathcal{K}}A^*(\theta)$, $\theta \in \mathfrak{R}^N$; i.e., $\text{cl}(\mathcal{E}) := \text{cl}\{Ax : x \in \mathcal{K}\} = \text{cl}(\text{range}(AP_{\mathcal{K}}A^*))$.*

Proof. [19] It must be shown that

$$\inf_{\theta \in \mathfrak{R}^N} \|Aq - AP_{\mathcal{K}}A^*(\theta)\|^2 = 0 \quad \forall q \in \mathcal{K}.$$

To this end, observe that $\ker(A)$ is orthogonal to $\text{range}(A^*) = \mathcal{G}$ and recall Π is the orthogonal projection onto \mathcal{G} . Thus, it must be shown that

$$\inf_{p \in \mathcal{G}} \|\Pi(q - P_{\mathcal{K}}(p))\|^2 = 0 \quad \forall q \in \mathcal{K}.$$

If $\mathcal{G} = \text{range}(A^*) = 0$, then $\Pi = 0$ and the claim is proven; so attention is restricted to the case of \mathcal{G} nontrivial.

Proceeding by contradiction, assume there exists some $q \in \mathcal{K}$ for which the infimum is $\epsilon > 0$. The closure of $\text{range}(\Pi P_{\mathcal{K}}\Pi)$ is convex from Lemma 4. Let z denote the nearest point in $\text{cl}(\text{range}(\Pi P_{\mathcal{K}}\Pi))$ to Πq , with $\|\Pi q - z\|^2 = \epsilon$. Then, there exists a hyperplane \mathcal{H} in the finite dimensional subspace \mathcal{G} containing $h = \frac{1}{2}(\Pi q + z)$ and normal to $(\Pi q - z)$. The hyperplane \mathcal{H} separates Πq from $\text{range}(\Pi P_{\mathcal{K}}\Pi)$.

Next, a point p_t is constructed to provide a contradiction. Let $p_t = \Pi q + t(\Pi q - z)$, $t > 0$. For t sufficiently large, p_t is closer to q than to \mathcal{H} . In particular, let $Q = \|(I - \Pi)q\|^2$ and observe

$$\|p_t - q\|^2 = \|\Pi q + t(\Pi q - z) - (\Pi q + (I - \Pi)q)\|^2 = t^2\epsilon + Q$$

On the other hand, the projection of p_t onto \mathcal{H} is h for all $t > 0$. Thus,

$$\inf_{y \in \mathcal{H}} \|p_t - y\|^2 = \|\Pi q + t(\Pi q - z) - \frac{1}{2}(\Pi q + z)\|^2 = \left(t + \frac{1}{2}\right)^2 \epsilon.$$

Hence, for $(t + \frac{1}{4})\epsilon > Q$, $d(p_t, q) < d(p_t, \mathcal{H})$, where $d(\cdot, \cdot)$ is adopted as a distance notation. Now, let $\mathcal{J} = \mathcal{H} \oplus \mathcal{G}^\perp$ and let \mathcal{J}^+ denote the halfspace in \mathcal{S} containing $\text{range}(\Pi P_{\mathcal{K}}\Pi)$. Then,

$$d(p_t, \mathcal{K}) \leq d(p_t, q) < d(p_t, \mathcal{H}) = d(p_t, \mathcal{J}) \leq d(p_t, \mathcal{J}^+ \cap \mathcal{K})$$

implying $P_{\mathcal{K}}(p_t) \in \mathcal{J}^-$, whence $\Pi P_{\mathcal{K}}(p_t) \in \mathcal{J}^-$ and $\Pi P_{\mathcal{K}}(p_t) \notin \text{range}(\Pi P_{\mathcal{K}} \Pi)$. But $\Pi p_t = p_t$, providing a contradiction. Therefore, it must be the case that the infimum is indeed zero. \square

The *relative interior* of a convex set $\mathcal{C} \subset \mathbb{R}^N$, denoted $\text{ri}(\mathcal{C})$, is defined as the interior that results when \mathcal{C} is regarded as a subset of the intersection of all closed linear varieties containing \mathcal{C} . Given convex sets \mathcal{C}_1 and \mathcal{C}_2 in \mathbb{R}^N , $\text{cl}(\mathcal{C}_1) = \text{cl}(\mathcal{C}_2)$ if and only if $\text{ri}(\mathcal{C}_1) = \text{ri}(\mathcal{C}_2)$ [23]. The desired existence result now follows immediately from the theorem. This result can also be developed from convex duality theory [4].

COROLLARY 2 (CONSTRAINT QUALIFICATION). *If $\beta \in \text{ri}(\mathcal{E})$, then there exists θ such that $AP_{\mathcal{K}}A^*(\theta) = \beta$, i.e., $\beta \in \text{range}(AP_{\mathcal{K}}A^*)$.*

Proof. From Theorem 2, $\text{cl}(\mathcal{E}) = \text{cl}(\text{range}(AP_{\mathcal{K}}A^*))$. Hence, equivalence of the relative interiors follows: $\text{ri}(\mathcal{E}) = \text{ri}(\text{range}(AP_{\mathcal{K}}A^*))$. \square

In an infinite-dimensional Hilbert space there exist closed convex sets \mathcal{K} without interior for which support points are only dense in the boundary and form only a set of the first category, the complementary set being dense as well [10]. Yet, a simple consequence of Corollary 2 is that for $\beta \in \text{ri}(\mathcal{E})$, the solution \hat{x} to (P) is, in fact, a support point of \mathcal{K} and, moreover, some normal to \mathcal{K} at \hat{x} intersects the subspace \mathcal{G} .

Two commonly employed but more restrictive constraint qualifications found in the literature follow as corollaries to the result in Theorem 2.

COROLLARY 3 (SLATER CONSTRAINT). *If \mathcal{K} has interior and the feasible set $\mathcal{F} := \mathcal{K} \cap \{x : Ax = \beta\}$ contains points interior to \mathcal{K} , then there exists $\hat{\theta}$ such that $AP_{\mathcal{K}}A^*(\hat{\theta}) = \beta$.*

COROLLARY 4 ([2], [8]). *Let $\mathcal{S} = L_2$ and let \mathcal{K} be the closed convex cone of nonnegative functions in L_2 . If $\beta \in \text{int}(\mathcal{E})$, then there exists $\hat{\theta}$ such that $AP_{\mathcal{K}}A^*(\hat{\theta}) = \beta$.*

Theorem 2 answers the question of existence of the parameterization $\hat{x} = P_{\mathcal{K}}A^*(\theta)$. The second issue, uniqueness of a parameter vector, is equivalent to the single-valuedness of the operator $f = (AP_{\mathcal{K}}A^*)^{-1}$.

PROPOSITION 1. *If \mathcal{E} has nonempty interior, then a parameter vector θ satisfying $AP_{\mathcal{K}}A^*(\theta) = \beta$ is unique for almost every $\beta \in \mathcal{E}$.*

Proof. The mapping $(AP_{\mathcal{K}}A^*)^{-1}$ is a monotone operator by Lemma 3. From [27, Thm. 1], the set of points where a monotone operator on a finite-dimensional Hilbert space is not single valued has zero Lebesgue measure. \square

Furthermore, the set of points in \mathcal{E} for which the representation is unique is a subset of the relative interior of $\text{range}(AP_{\mathcal{K}}A^*)$ [22, Cor. 1.1]. For linearly dependent measurement signals $\{g_1, \dots, g_N\}$ the extendible set $\mathcal{E} \subset \mathbb{R}^N$ is contained in a subspace of dimension less than N and $\text{int}(\mathcal{E}) = \emptyset$.

5. Iterative computation. From Theorems 1 and 2, the solution \hat{x} to (P) is parameterized by $\hat{x} = P_{\mathcal{K}}A^*(\hat{\theta})$, where the vector $\hat{\theta}$ solves the nonlinear system $AP_{\mathcal{K}}A^*(\theta) = \beta$. Equivalently, the parameter vector $\hat{\theta}$ is a fixed point of the operator $T : \mathbb{R}^N \mapsto \mathbb{R}^N$ defined by $T(\theta) = \theta + \beta - AP_{\mathcal{K}}A^*(\theta)$. The operator T is not a contraction, nor does it have a compact domain; therefore, the well-known Banach and Brouwer fixed point results are not applicable. Nonetheless, the properties of firmly nonexpansive and averaged mappings are exploited to show that the sequence of Picard iterations

$$(1) \quad \theta^{(n+1)} = \theta^{(n)} + \lambda[\beta - AP_{\mathcal{K}}A^*(\theta^{(n)})], \quad \lambda \in (0, 2)$$

converges to a fixed point of T . Additionally, the sequence is shown to characterize a least-squares solution to (P) when there exists no fixed point.

Case 1. First, the sequence $\{\theta^{(n)}\}$ is considered for the case in which T has a fixed point. Convergence is established by relying on three simple lemmas in Euclidean N -space. A mapping $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ is said to be *nonexpansive* if $\|f(x) - f(y)\| \leq \|x - y\|$ for all x, y in \mathbb{R}^N . Further, f is *firmly nonexpansive* if and only if $2f - I$ is nonexpansive [7].

LEMMA 5. *As defined above, let T be the operator given by $T(\theta) = \theta + \beta - AP_K A^*(\theta)$. If the measurement signals $\{g_1, \dots, g_N\}$ satisfy $\sum_{k=1}^N \|g_k\|^2 \leq 1$, then T is firmly nonexpansive.*

Proof. To show that T is firmly nonexpansive, $2T - I$ is shown to be nonexpansive. To this end, direct computation using the definitions of T and of the adjoint A^* yields

$$\|(2T - I)x - (2T - I)y\|^2 \leq \|x - y\|^2 \Leftrightarrow \langle P_K(x') - P_K(y'), x' - y' \rangle \geq \|A(P_K(x') - P_K(y'))\|^2,$$

where $x' := A^*x$ and $y' := A^*y$. From Lemma 2, $\langle P_K(x') - P_K(y'), x' - y' \rangle \geq \|P_K(x') - P_K(y')\|^2$. Furthermore, by hypothesis on the measurement signals and application of the Cauchy-Bunyakovskii-Schwarz inequality, A is nonexpansive:

$$\|Aw\|^2 = \sum_{k=1}^N (\langle g_k, w \rangle)^2 \leq \sum_{k=1}^N \|g_k\|^2 \|w\|^2 \leq \|w\|^2 \quad \forall w \in \mathcal{S}.$$

Hence,

$$\|P_K(x') - P_K(y')\|^2 \geq \|A(P_K(x') - P_K(y'))\|^2,$$

and T is firmly nonexpansive. \square

LEMMA 6 ([7]). *Let $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ be a nonexpansive operator with a fixed point. Then, $\{f^n(x)\}$ converges to a fixed point of f if and only if f is asymptotically regular, i.e.,*

$$\lim_n f^n(x) - f^{n+1}(x) = 0 \text{ for all } x \in \mathbb{R}^N.$$

As an example of a nonexpansive operator \mathbb{R} to \mathbb{R} with a fixed point and not asymptotically regular, consider $f(x) = -x - 1$. Although a nonexpansive operator f may not be asymptotically regular, the *averaged mapping* $f_\lambda := \lambda f + (1 - \lambda)I$, where $0 < \lambda < 1$, shares the same fixed point set and has desirable asymptotic properties.

LEMMA 7 ([5]). *Let f be a nonexpansive operator in \mathbb{R}^N . Although the operator f itself may not be asymptotically regular, if f has a fixed point, then the averaged mapping f_λ is asymptotically regular.*

The result now follows directly.

THEOREM 3. *Assume T has a fixed point. For $\lambda \in (0, 2)$ let $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ be defined by*

$$f(\theta) = \theta + \lambda[\beta - AP_K A^*(\theta)]$$

If the measurement signals $\{g_1, \dots, g_N\}$ satisfy $\sum_{k=1}^N \|g_k\|^2 \leq 1$, then the sequence of Picard iterates $\{f^n(\theta)\}$ converges to a fixed point of T for any $\theta \in \mathbb{R}^N$.

Proof. From Lemma 5, T is firmly nonexpansive, so $2T - I$ is nonexpansive and has the same fixed point set as T . Simply note that for $0 < \delta < 1$, f is the averaged mapping $f = \delta(2T - I) + (1 - \delta)I = 2\delta T + (1 - 2\delta)I$. By Lemma 7, f is asymptotically regular. Application of Lemma 6 then yields Picard iterates $\{f^n(\theta)\}$ converging to a

fixed point of $2T - I$. Thus, the limit $\hat{\theta}$ is a fixed point of T and, therefore, satisfies $AP_{\mathcal{K}}A^*(\hat{\theta}) = \beta$. \square

For all β in the relative interior of the extendible set \mathcal{E} , T has a fixed point by Corollary 2, and the Picard iteration $\{f^n(\theta)\}$ yields the solution to (P). The hypothesis that the measurement signals have square sum not exceeding one can always be satisfied by simple scaling.

Case 2. Next, the behavior of the sequence $\{f^n(\theta)\}$ is considered for the general case in which T may be fixed point free. First, T is trivially fixed point free when the measurement vector β is not extendible, i.e., when there exists no signal x in the constraint set \mathcal{K} for which $Ax = \beta$, and hence, no solution to (P). In application, such a nonextendible vector β may result from either measurement noise or from failure of the constraint set \mathcal{K} to reflect physical reality. In addition, T may have no fixed point for β in the relative boundary of \mathcal{E} .

The objective of determining a feasible signal $x \in \mathcal{S}$ satisfying both $x \in \mathcal{K}$ and $Ax = \beta$ is unobtainable when β fails to lie in the extendible set. A well-motivated and popular recourse is to find a signal in the constraint set \mathcal{K} that best matches the measurement vector β in the least-squares sense: $\inf_{x \in \mathcal{K}} \|Ax - \beta\|$. (This choice implicitly supposes greater confidence in the knowledge expressed by the constraint set \mathcal{K} than in the noisy measurement β .) If more than one signal achieves this infimum, then the unique infimizer of minimum norm is termed the *minimum norm least-squares* solution and solves

$$(P') \quad \min_{x \in \mathcal{K}} \|x\| \quad \text{subject to} \quad \|Ax - \beta\| = \inf_{y \in \mathcal{K}} \|Ay - \beta\|.$$

A weighted least-squares formulation is easily adopted with corresponding change in the definition of the adjoint, A^* . For a closed convex constraint set \mathcal{K} and a linear measurement operator A , the extendible set $\mathcal{E} = A(\mathcal{K}) \subset \mathbb{R}^N$, though not necessarily closed, is convex. Hence, for a measurement vector $\beta \notin \mathcal{E}$, there exists a unique closest vector in the closure of \mathcal{E} , namely, the projection of β onto $cl(\mathcal{E})$, $P_{\bar{\mathcal{E}}}(\beta)$.

PROPOSITION 2. *The infimum in (P') is achieved if and only if $P_{\bar{\mathcal{E}}}(\beta)$ is in \mathcal{E} .*

Proof. With $P_{\bar{\mathcal{E}}}$ as above, $\inf_{y \in \mathcal{K}} \|Ay - \beta\| = \|P_{\bar{\mathcal{E}}}(\beta) - \beta\|$. \square

Therefore, for $P_{\bar{\mathcal{E}}}(\beta) \in \mathcal{E}$, problem (P') is equivalent to (P) with measurement vector $P_{\bar{\mathcal{E}}}(\beta)$.

The asymptotic behavior of averaged mappings provides the solution to (P'), as readily demonstrated using the following asymptotic property of nonexpansive maps.

LEMMA 8 ([1]). *Let $h : \mathbb{R}^N \mapsto \mathbb{R}^N$ be nonexpansive and define the averaged mapping $h_{\lambda} = \lambda h + (1 - \lambda)I$, $\lambda \in (0, 1)$. Then for all θ in \mathbb{R}^N*

$$(a) \quad \lim_{n \rightarrow \infty} \frac{h_{\lambda}^n(\theta)}{n} = -\nu$$

$$(b) \quad \lim_{n \rightarrow \infty} [h_{\lambda}^n(\theta) - h_{\lambda}^{n+1}(\theta)] = \nu,$$

where ν is the unique point of least norm in $cl(\text{range}(I - h_{\lambda}))$. Additionally, h has no fixed point if and only if $\lim_{n \rightarrow \infty} \|h_{\lambda}^n(\theta)\| = \infty$ for all θ in \mathbb{R}^N .

In relation to the asymptotic regularity condition of Lemma 6, observe that $h^n(\theta)$ is a Cauchy sequence if and only if h has a fixed point and ν is the zero vector.

THEOREM 4. *Let $\beta \in \mathbb{R}^N$ be an observed measurement vector, and let $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ be defined by $f(\theta) = \theta + \lambda[\beta - AP_{\mathcal{K}}A^*(\theta)]$ for $\lambda \in (0, 2)$. Also, assume $\sum_{k=1}^N \|g_k\|^2 \leq 1$. Let $\{\theta^{(n)}\}$ denote the sequence of Picard iterates $\theta^{(n)} = f^n(\theta^{(0)})$*

with initial iterate $\theta^{(0)}$. Then, for any $\theta^{(0)} \in \mathfrak{R}^N$, the sequence $\{AP_{\mathcal{K}}A^*(\theta^{(n)})\}$ converges to $P_{\mathcal{E}}(\beta)$, the projection of β onto the closure of the extendible set.

Proof. [19] By Theorem 2, $\text{cl}(\text{range}(AP_{\mathcal{K}}A^*)) = \text{cl}(\mathcal{E})$. Therefore, the closure of the range of $(I - f)$ is simply a scaled translate of the closure of the extendible set:

$$\text{cl}(\text{range}(I - f)) = \text{cl}\{\gamma : \gamma = \lambda(AP_{\mathcal{K}}A^*(\theta) - \beta), \theta \in \mathfrak{R}^N\} = \lambda\{\text{cl}(\mathcal{E}) - \beta\}.$$

Then, the minimum norm element of $\text{cl}(\text{range}(I - f))$ is $\lambda\nu$, where ν is the minimum norm element of $\text{cl}(\mathcal{E}) - \beta$. Hence, the projection of β onto $\text{cl}(\mathcal{E})$ is given by the sum $P_{\mathcal{E}}(\beta) = \beta + \nu$. By Lemma 8, given $\epsilon > 0$, there exists some integer M_{ϵ} such that for all n exceeding M_{ϵ}

$$\begin{aligned} \epsilon &> \|\lambda\nu - (\theta^{(n)} - \theta^{(n+1)})\| \\ &= \|\lambda\nu - \theta^{(n)} + \theta^{(n)} + \lambda[\beta - AP_{\mathcal{K}}A^*(\theta^{(n)})]\| \\ &= \lambda\|P_{\mathcal{E}}(\beta) - AP_{\mathcal{K}}A^*(\theta^{(n)})\|. \end{aligned}$$

Hence, $AP_{\mathcal{K}}A^*(\theta^{(n)}) \rightarrow P_{\mathcal{E}}(\beta)$ as $n \rightarrow \infty$. \square

COROLLARY 5. *If, in addition to the hypotheses of Theorem 4, $P_{\mathcal{E}}(\beta)$ is contained in the extendible set \mathcal{E} , then the sequence of approximate reconstructions $\{x^{(n)}\}$, $x^{(n)} := P_{\mathcal{K}}A^*(\theta^{(n)})$, is bounded, and there exists a subsequence $\{x^{(n_j)}\}$ that converges weakly to \hat{x} , the solution to (P') .*

Proof. From Proposition 2, there exists a solution, \hat{x} , to (P') . By Theorem 4, $\epsilon_n := \|\Pi(\hat{x} - x^{(n)})\| \rightarrow 0$. Then, employ Lemma 2, a direct sum decomposition with \mathcal{G} , and the Cauchy-Bunyakovskiĭ-Schwarz inequality to learn

$$\begin{aligned} 0 &\leq \|\hat{x} - A^*\theta^{(n)}\| - \|P_{\mathcal{K}}A^*\theta^{(n)} - A^*\theta^{(n)}\| \\ &\leq \|(I - \Pi)(\hat{x})\| - \|(I - \Pi)(P_{\mathcal{K}}A^*\theta^{(n)})\| + \\ &\quad \left| \|\Pi(\hat{x} - A^*\theta^{(n)})\| - \|\Pi(P_{\mathcal{K}}A^*\theta^{(n)} - A^*\theta^{(n)})\| \right| \\ &\leq \|(I - \Pi)(\hat{x})\| - \|(I - \Pi)(P_{\mathcal{K}}A^*\theta^{(n)})\| + \epsilon_n. \end{aligned}$$

Therefore, $x^{(n)} = P_{\mathcal{K}}A^*\theta^{(n)}$ is a bounded sequence in \mathcal{K} , and consequently there exists some subsequence $\{x^{(n_j)}\}$ that converges weakly. Let y be the weak limit. Now, $y \in \mathcal{K}$, $\Pi y = \Pi\hat{x}$, and $Ay = P_{\mathcal{E}}(\beta)$ by Theorem 4. Thus, $\|\hat{x}\| \leq \|y\|$ by definition of \hat{x} . Conversely, $\|(I - \Pi)(y)\| \leq \|(I - \Pi)(\hat{x})\|$ from above, whence $\|y\| = \|\hat{x}\|$. Since, from Lemma 1, \hat{x} is the unique element of minimum norm in \mathcal{K} for which $A\hat{x} = P_{\mathcal{E}}(\beta)$, it follows that $y = \hat{x}$. \square

A practical criterion for convergence in computer implementation of Theorem 4 is to test the sequence of differences in successive iterations for convergence to $\lambda\nu$ within a given tolerance. However, the vector ν is not known a priori and is zero if and only if β is an observation vector in the closure of the extendible set. Nonetheless, $\{\theta^{(n)} - \theta^{(n+1)}\}$ is indeed a Cauchy sequence in \mathfrak{R}^N by Lemma 8. Therefore, observing that

$$\theta^{(n)} - \theta^{(n+1)} = \theta^{(n)} - f(\theta^{(n)}) = \lambda[P_{\mathcal{K}}A^*(\theta^{(n)}) - \beta]$$

is simply the residual error scaled by λ , the iterations may be terminated when the change in the residual error from iterate n to $n + 1$ is less than some prescribed value. Moreover, this sequence of residual errors is monotonically nonincreasing in norm due to the nonexpansiveness of f . Although $\{\theta^{(n)}\}$ is divergent, it grows only linearly as $n\lambda\nu$. Therefore, the divergence presents no practical computational overflow problems, even for a large number of iterations, since $\|\nu\|^2$ is bounded by the noise power in the measurement β .

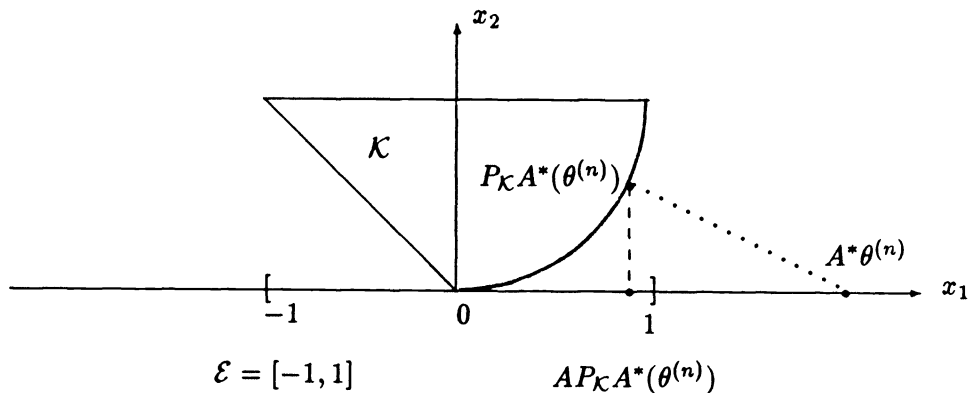


FIG. 2. An example in the plane.

6. Example. The results of §§3, 4, and 5 are illustrated by a simple example in the Euclidean plane. Although analytical nuances are lost from the infinite-dimensional case, the relationships among the relative interior of the extendible set, fixed points, and Picard iterations are clearly illuminated. (An example application to an infinite-dimensional problem is found in [21].) In the Hilbert space $\mathcal{S} = \mathbb{R}^2$ let the constraint set \mathcal{K} be the closed convex set depicted in Fig. 2. Let the measurement β be simply the first coordinate of a vector in \mathbb{R}^2 . Accordingly, the single measurement signal is $g_1 = [1 \ 0]^t$, yielding $A : \mathbb{R}^2 \mapsto \mathbb{R}$ given by $[1 \ 0]$ and $A^* = g_1$. The extendible set $\mathcal{E} := A(\mathcal{K})$ is the closed interval $[-1, 1]$. That \mathcal{E} is closed is implied by the boundedness of \mathcal{K} . For a given β , the feasible set \mathcal{F} is the intersection of \mathcal{K} with the line $x_1 = \beta$.

By Corollary 2, if the measurement is in the open interval $\text{ri}(\mathcal{E}) = (-1, 1)$, then there exists some scalar θ such that $\hat{x} = P_{\mathcal{K}} A^*(\theta)$ is the solution to (P). For the measurement $\beta = 1$ on the boundary of \mathcal{E} , no finite θ provides a parameterization; the measurement $\beta = -1$ is likewise on the boundary of \mathcal{E} , yet $\theta = -2$ provides the solution to (P). The dense uniqueness in Proposition 1 is illustrated by the infinitely many parameterizations, $\theta \in (\infty, -2]$, for $\beta = -1$. (A translation of \mathcal{K} by $[0 \ 1]^t$ provides an example of nonunique parameterization for β in the interior of \mathcal{E} .)

The iterative procedure of Theorems 3 and 4 is given by

$$f^{n+1}(\theta^{(0)}) = \theta^{(n+1)} = \theta^{(n)} + \lambda[\beta - AP_{\mathcal{K}} A^*(\theta^{(n)})], \quad \lambda \in (0, 2).$$

The action of $AP_{\mathcal{K}} A^* : \mathbb{R} \mapsto \mathbb{R}$ is depicted in Fig. 2 and is given by

$$AP_{\mathcal{K}} A^*(\theta) = \begin{cases} -1, & \theta \leq -2 \\ \frac{1}{2}\theta, & -2 < \theta < 0 \\ \theta(\theta^2 + 1)^{-\frac{1}{2}}, & \theta \geq 0. \end{cases}$$

The existence of a parameterization θ is equivalent to the existence of a fixed point for f . For $\beta \in [-1, 1)$, there exists a fixed point for f , and by Theorem 3, the sequence $\{\theta^{(n)}\}$ converges to a parameter yielding \hat{x} , the minimum norm element of the feasible set. For $\beta \geq 1$, f has no fixed point and $\{|\theta^{(n)}|\}$ diverges as $\lambda n \nu$, where $\nu = \beta - 1$ is the distance of β from the extendible set. Yet, by Theorem 4, the sequence $\{AP_{\mathcal{K}} A^*(\theta^{(n)})\}$ converges to $P_{\bar{\mathcal{E}}}$, and by Corollary 5, a subsequence of the

approximate reconstructions $\{P_{\mathcal{K}}A^*(\theta^{(n)})\}$ converges to the minimum norm, least-squares solution to (P') , $\hat{x} = [1 \ 1]^t$. Finally, the parameter vector is unique for every β in \mathcal{E} except $\beta = -1$, where the solution to (P) is not a regular point of \mathcal{K} .

7. Discussion. The method of successive projections is an alternative scheme for computing an element in the feasible set [26]. Treating the variety $Ax = \beta$ as an additional closed convex constraint set \mathcal{K}_{M+1} , the iteration

$$(2) \quad x^{(n+1)} = (P_{\mathcal{K}_{M+1}}P_{\mathcal{K}_M}P_{\mathcal{K}_{M-1}} \cdots P_{\mathcal{K}_1})x^{(n)}$$

converges weakly in \mathcal{S} to an element of the feasible set, provided one exists. The method is attractive in that any number of convex constraint sets may be incorporated without requiring synthesis of the projection onto the intersection $\mathcal{K} = \bigcap_{j=1}^M \mathcal{K}_j$. However, the technique does not allow the preferential selection of one feasible signal over others, as provided by the optimality criterion in (P') . In general, the limit point of Eq. (2) depends on both the initial estimate $x^{(0)}$ and the ordering of the composition of projection operators. Moreover, the iterations are performed in the (perhaps infinite-dimensional) signal space \mathcal{S} rather than in \mathfrak{R}^N and typically suffer from slow convergence rates and high computational cost per iteration [11], [24]. In addition, successive projections do not in general provide a least-squares solution when no feasible signal exists.

In contrast, the signal recovery algorithm established in Theorems 1–4 provides a finite-dimensional parameterization for a signal reconstruction. The iterative algorithm is performed in the parameter space to preferentially produce the unique least-squares solution consistent with the constraints and closest to a specified nominal signal. Moreover, Newton-Raphson iterations may typically be applied in \mathfrak{R}^N for quadratically convergent iterative computation; the requisite derivatives are guaranteed to exist almost everywhere since $AP_{\mathcal{K}}A^*$ is Lipschitz. A potential difficulty in implementing the iterative scheme in Eq. (1) is the need to construct $AP_{\mathcal{K}}A^*$, which may require numerical approximation of the projection onto \mathcal{K} , the intersection of constraint sets. Yet, in application, \mathcal{K} is physically motivated and, as such, typically gives rise to an intuitive and tractable projection operator. Furthermore, sensitivity of the solution \hat{x} to errors in the parameters $\hat{\theta}$ is low since $P_{\mathcal{K}}A^*$ is nonexpansive. Although the constraint set must be convex and the signal space is Hilbertian, the formulation admits a large and relevant class of sets for incorporating prior information.

Many well-known linear reconstruction results follow immediately from Theorem 1 for the special case of constraint sets \mathcal{K} that are subspaces, e.g., [3], [6], [12], [16]. Likewise, Theorems 1 and 2 extend, in Hilbert space, the optimization results in [2], [8], and [9] from closed convex cones to arbitrary closed convex constraint sets. For example, the minimum energy correlation extension presented in [8] and [25] may be directly obtained with $\mathcal{S} = L_2$ and \mathcal{K} the convex cone of nonnegative spectral estimates. Kuhn-Tucker, Lagrange multiplier, and Fenchel duality theorems, e.g., [13] are similar dual optimization results that have been applied to signal recovery problems and, in addition, admit cost functions more general than the weighted norm. However, the hypotheses of these classical results require nonempty interior and regularity conditions that are absent in Theorems 1–4. These seemingly technical restrictions are, in fact, of great importance to application in many practical reconstruction tasks since, for example, the set of nonnegative signals in L_2 is without interior.

8. Conclusion. Motivated by practical reconstruction, estimation, and interpolation problems, an explicit solution to a constrained minimization problem has been

derived. The finite parameterization led to a simple and computationally attractive iterative algorithm. The constraint qualification for the infinite-dimensional program with linear equality constraints and a convex set constraint extended previous results for a convex cone set constraint.

Acknowledgment. The authors gratefully acknowledge the helpful, insightful criticisms of an anonymous reviewer.

REFERENCES

- [1] J. B. BAILLON, R. E. BRUCK, AND S. REICH, *On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces*, Houston J. Math., 4 (1978), pp. 1–9.
- [2] A. BEN-TAL, J. M. BORWEIN, AND M. TEBoulLE, *A dual approach to multidimensional L_p spectral estimation problems*, SIAM J. Control Optim., 26 (1988), pp. 985–996.
- [3] M. BERTERO, C. DEMOL, AND E. R. PIKE, *Linear inverse problems with discrete data I: General formulation and singular system analysis*, Inverse Problems, 1 (1985), pp. 301–330.
- [4] J. M. BORWEIN AND A. LEWIS, *Partially finite convex programming, Part I: Quasi relative interiors and duality theory*, Mathematical Programming, to appear.
- [5] F. E. BROWDER AND W. V. PETRYSHYN, *The solution by iteration of nonlinear functional equations in Banach spaces*, Bull. Amer. Math. Soc., 72 (1966), pp. 571–575.
- [6] C. L. BYRNE AND R. M. FITZGERALD, *Spectral estimators that extend the maximum entropy and maximum likelihood methods*, SIAM J. Appl. Math., 44 (1984), pp. 425–442.
- [7] K. GOEBEL AND S. REICH, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, Marcel Dekker, New York, 1984.
- [8] R. K. GOODRICH AND A. STEINHARDT, *L_2 spectral estimation*, SIAM J. Appl. Math., 46 (1986), pp. 417–426.
- [9] L. D. IRVINE AND P. W. SMITH, *Constrained minimization in a dual space*, in Methods of Functional Analysis in Approximation Theory, C. A. Micchelli, D. V. Pai, and B. V. Limaye, eds., Bombay, 1986, Birkhauser-Verlag, pp. 205–219.
- [10] V. L. KLEE, *The support property of a convex set in a linear normed space*, Duke Math. J., 15 (1948), pp. 767–772.
- [11] R. M. LEAHY AND C. E. GOUTIS, *An optimal technique for constraint-based image restoration and reconstruction*, IEEE Trans. Acoust. Speech Signal Process., ASSP-34 (1986), pp. 1629–1642.
- [12] L. LEVI, *Fitting a band-limited signal to given points*, IEEE Trans. Inform. Theory, IT-11 (1965), pp. 372–376.
- [13] D. G. LUENBERGER, *Optimization by Vector Space Methods*, Wiley, New York, 1969.
- [14] B. P. MEDOFF, W. R. BRODY, AND A. MACOVSKI, *The use of a priori information in image reconstruction from limited data*, in Proc. 1983 IEEE Intl. Conf. Acoust. Speech Signal Process., Boston, 1983, pp. 131–134.
- [15] G. J. MINTY, *On the maximal domain of a monotone function*, Michigan Math. J., 8 (1961), pp. 135–137.
- [16] A. PAPOULIS, *A new algorithm in spectral analysis and band-limited extrapolation*, IEEE Trans. Circuits and Systems, CAS-22 (1975), pp. 735–742.
- [17] A. PAZY, *Asymptotic behavior of contractions in Hilbert space*, Israel J. Math., 9 (1971), pp. 235–240.
- [18] L. C. POTTER, *A finite parameterization for constrained minimum norm interpolants*, in Proc. 32nd Midwest Symposium on Circuits and Systems, Urbana, IL, 1989, pp. 1174–1177.
- [19] ———, *Constrained Signal Reconstruction*, Ph.D. thesis, University of Illinois, Urbana, 1990.
- [20] L. C. POTTER AND K. S. ARUN, *An iterative algorithm for minimum-norm, constrained extrapolation*, in Proc. 22nd Asilomar Confer. Signals Systems Comput., Pacific Grove, CA, November 1988.
- [21] L. C. POTTER, K. S. ARUN, AND D. L. JONES, *Recovery of pore-size distributions from NMR spin-lattice relaxation measurements*, in Proc. 1990 Digital Signal Processing Workshop, New Paltz, NY, September 1990, pp. 2.1.1–2.1.2.
- [22] R. T. ROCKAFELLAR, *Local boundedness of nonlinear, monotone operators*, Michigan Math. J., 16 (1969), pp. 397–407.
- [23] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

- [24] M. I. SEZAN AND H. STARK, *Image restoration by the method of convex projections: Part 2—Applications and numerical results*, IEEE Trans. Med. Imaging, MI-1 (1982), pp. 95–101.
- [25] A. O. STEINHARDT, R. K. GOODRICH, AND R. A. ROBERTS, *Spectral estimation via minimum energy correlation extension*, IEEE Trans. Acoust. Speech Signal Process., ASSP-33 (1985), pp. 1509–1515.
- [26] D. C. YOULA AND H. WEBB, *Image restoration by the method of convex projections: Part 1—Theory*, IEEE Trans. Med. Imaging, MI-1 (1982), pp. 81–94.
- [27] E. H. ZARANTONELLO, *Dense single-valuedness of monotone operators*, Israel J. Math., 15 (1973), pp. 158–166.

WHITE-NOISE REPRESENTATIONS IN STOCHASTIC REALIZATION THEORY*

VIVEK S. BORKAR†

Abstract. It is proved that any random sequence can be exhibited as the output of a stochastic dynamical system driven by white noise. Further refinements are obtained for Markov, stationary, and ergodic sequences. This settles some open problems in stochastic realization theory posed by Willems and Van Schuppen [*NATO ASI-AMS Seminar on Algebraic and Geometric Methods in Linear System Theory*, Harvard University, Cambridge, MA, 1979].

Key words. white-noise representation, stochastic realization theory, stochastic dynamical systems, ergodic decomposition of stationary processes, extremal measures

AMS subject classifications. 93E03, 60G05, 60G10

1. Introduction. The white-noise representation problem in stochastic realization theory is to show a given random sequence as the output of a white-noise-driven stochastic dynamical system. (This is made precise later.) The goal of this paper is to prove some conjectures of Willems and Van Schuppen [10] in this context. The paper is organized as follows: The remainder of this section recalls the problems of stochastic realization theory from [10]. Section 2 contains derivations of results concerning extremal probability measures on a product of Polish spaces with a given marginal. Section 2 somewhat overlaps [2], where similar results are used in a different context. In §3 the main result and some consequences thereof are proved. Section 4 explores the connection between the ergodic decomposition of stationary processes and their white-noise representation.

The results proved here were conjectured in [10]. See [9] and [10] for extensive surveys of stochastic realization theory and further references, and see [4] for the important special case of linear stochastic realization theory.

The principal problems of abstract stochastic realization theory are as follows.

(i) *Strong stochastic realization problem:* Given a random sequence $\{Y_n\}$ on a probability space (Ω, \mathcal{F}, P) , construct on this probability space another sequence of random variables $\{X_n\}$ such that $[(X_{n-1}, Y_{n-1}), (X_{n-2}, Y_{n-2}), \dots]$ and $[(X_n, Y_n), (X_{n+1}, Y_{n+1}), \dots]$ are conditionally independent, given X_n for each n . (In particular, $\{X_n\}$ is Markov.)

(ii) *Weak stochastic realization problem:* Given $\{Y_n\}$ as in problem (i), construct on some probability space $(\Omega', \mathcal{F}', P')$ random sequences $\{X'_n\}$ and $\{Y'_n\}$ such that $[(X'_{n-1}, Y'_{n-1}), (X'_{n-2}, Y'_{n-2}), \dots]$ and $[(X'_n, Y'_n), (X'_{n+1}, Y'_{n+1}), \dots]$ are conditionally independent given X'_n for each n and that $\{Y_n\}$ and $\{Y'_n\}$ agree in law. (Again, $\{X'_n\}$ is Markov.)

(iii) *Strong white-noise representation problem:* Given a strong realization as in (i), construct on the same probability space two independent and identically distributed sequences $\{W_n\}$, $\{W'_n\}$ independent of each other such that

$$(1.1) \quad X_{n+1} = f_n(W_n, X_n),$$

$$(1.2) \quad Y_n = g_n(W'_n, X_n)$$

for each n and measurable $\{f_n\}$ and $\{g_n\}$. $\{W_n\}$ and $\{W'_n\}$ are called white-noise sequences (signal and observation noise, respectively).

* Received by the editors May 20, 1991; accepted for publication (in revised form) February 6, 1992.

† Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India.

(iv) *Weak white-noise representation problem*: This is the same as problem (ii) for a given weak realization, except that $\{W_n\}$, $\{W'_n\}$ are now constructed on $(\Omega', \mathcal{F}', P')$ or an augmentation of it. (Here and later, an augmentation of a probability space (Ω, \mathcal{F}, P) will mean a new probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$ such that for some measurable space (E, ξ) , $\bar{\Omega} = \Omega \times E$, $\bar{\mathcal{F}} = \mathcal{F} \times \xi$, and P is the image of \bar{P} under the projection $(\bar{\Omega}, \bar{\mathcal{F}}) \rightarrow (\Omega, \mathcal{F})$.)

(v) *Minimal realization problem*: A strong (respectively, weak) realization as in problem (i) (respectively, problem (ii)) is minimal if there is no other strong (respectively, weak) realization $\{\tilde{X}_n\}$ and $\{Y_n\}$ (respectively, $\{\tilde{X}_n\}$, and $\{Y'_n\}$) on the same probability space such that $\sigma(\tilde{X}_i, i \leq n) \subset \sigma(X_i, i \leq n)$ (respectively, $\sigma(X'_i, i \leq n)$) for each n with the inclusion being strict for at least one n . Characterize all minimal realizations.

We settle problem (iv) above completely and prove by example that problem (iii) does not always have a solution.

We have not yet specified the hypotheses on the spaces in which the above sequences take values. We assume these to be Polish spaces, i.e., separable and metrizable with a complete metric. An important result concerning Polish spaces is that, given two Polish spaces of the same cardinality (in particular, two uncountable Polish spaces), there is a measurable isomorphism between them ([5, Thm. 2.12]). We use this result often and refer to it simply as the *isomorphism theorem*. For any Polish space X , $P(X)$ will denote the Polish space of probability measures on X with Prohorov topology [1], [5].

2. Extremal measures with a given marginal. In this section, we recall from [2] some results on probability measures on a product of Polish spaces having a given marginal on one of them. Details are included to make the present account self-contained.

Let S_1, S_2 be Polish spaces endowed with their Borel σ -fields. Let μ be a probability measure on $S_1 \times S_2$ that disintegrates [8] as

$$(2.1) \quad \mu(dx, dy) = \nu(dx)v(x, dy),$$

where ν is the image of μ under the projection $S_1 \times S_2 \rightarrow S_1$ and $x \rightarrow v(x, \cdot) : S_1 \rightarrow P(S_2)$ is the regular conditional law defined ν -almost surely uniquely. Let Q be the set of all such μ when ν is a prescribed element of $P(S_1)$. Q is clearly closed convex in $P(S_1 \times S_2)$. Let Q_e be the set of all extreme points of Q , and let $Q_D \subset Q$ be the set of those μ for which $v(x, \cdot)$ is a Dirac measure for ν -almost surely x . The main result of this section is that $Q_e = Q_D$.

Let $f \in C_b(S_2)$, and let $q \in P(S_2)$. Let $b \in R$ be the unique least number such that

$$g(\{x | f(x) < b\}) \leq \frac{1}{2},$$

$$q(\{x | f(x) > b\}) \leq \frac{1}{2}.$$

Let $A_1 = \{x | f(x) < b\}$, $A_2 = \{x | f(x) > b\}$, and $A_3 = \{x | f(x) = b\}$. Let $\delta \in [0, 1]$ be such that

$$q(A_1) + \delta q(A_3) = q(A_2) + (1 - \delta)q(A_3) = \frac{1}{2},$$

where $\delta = 0$ when $q(A_3) = 0$. Define $\alpha_1(q), \alpha_2(q) \in P(S_2)$ by

$$\alpha_1(q) = 2(I_{A_1} + \delta I_{A_3})q,$$

$$\alpha_2(q) = 2(I_{A_2} + (1 - \delta)I_{A_3})q.$$

It is not difficult to verify that α_1, α_2 are measurable maps $P(S_2) \rightarrow P(S_2)$. In fact, by [3, Thm. 2.1] this is equivalent to verifying that, for all Borel $A \subset S_2$, the map $q \rightarrow \alpha_1(q)(A)$ is measurable for $i = 1, 2$. Consider, say, $i = 1$. Then

$$\alpha_1(q)(A) = 2q(A \cap A_1) + 2\delta q(A \cap A_3),$$

where A_1 , A_3 , and δ depend on q through their dependence on b . Write b as $b(q)$ to make its q dependence explicit. By the foregoing and [3, Thm. 2.1], it suffices to prove that the map $q \rightarrow b(q)$ is measurable. However, for $c \in R$,

$$\{q | b(q) > c\} = \{q | q(\{x | f(x) \geq c\}) > \frac{1}{2}\},$$

which is measurable by [3, Thm. 2.1], and we are done.

Note that

$$(2.2) \quad q = \frac{1}{2}(\alpha_1(q) + \alpha_2(q)).$$

LEMMA 2.1. *If $\mu \in Q \setminus Q_D$, there exist a Borel set $A \subset S_1$ and an $f \in C_b(S_2)$ such that $\nu(A) > 0$, and, for $x \in A$, $f(\cdot)$ is not a constant $\nu(x)$ -almost surely.*

Proof. Let $\{f_i\}$ be a countable subset of $C_b(S_2)$ that separates points of $P(S_2)$. (See Remark 2.1 below.) Suppose that for ν -almost surely x , f_i is $\nu(x)$ -almost surely a constant for all i . Then $\nu(x)$ is a Dirac measure for such x , contradicting the hypothesis $\mu \notin Q_D$. Thus there exists a Borel set $A' \subset S_1$ such that $\nu(A') > 0$ and for $x \in A'$, f_i is not a constant $\nu(x)$ -almost surely for some i . Let

$$A_i = \{x \in S_1 | f_i \text{ is not a constant } \nu(x)\text{-almost surely}\}, \quad i = 1, 2.$$

Then A_i is the complement of

$$(2.3) \quad \bigcap_m \left\{x \in S_1 \mid \int g_m f_i \, d\nu(x) = \int g_m \, d\nu(x) \int f_i \, d\nu(x)\right\}$$

for a countable collection $\{g_m\}$ in $C_b(S_2)$ with the property that $\int g_i \, dm_1 = \int g_i \, dm_2$ for all i implies that $m_1 = m_2$ for finite signed measures m_1 and m_2 on S_2 . Expression (2.3) and hence A_i are measurable. We may set $A' = \bigcup_i A_i$. Then $\nu(A') > 0$ implies that $\nu(A_{i_0}) > 0$ for some i_0 , and the claim follows with $f = f_{i_0}$, $A = A_{i_0}$. \square

Remark 2.1. The family $\{f_i\}$ can be chosen as follows. Let d be any complete metric on S_2 taking values in $[0, 1]$ and consistent with the topology of S_2 . Map S_2 homeomorphically onto a G_δ subset S'_2 of $[0, 1]^\infty$ as in [1, pp. 219–220]. Then S'_2 is compact, and therefore $C(\bar{S}'_2)$ is separable. Take a countable dense subset of $C(\bar{S}'_2)$, restrict it to S'_2 , and pull it back by means of the homeomorphism.

A similar remark applies to the choice of $\{g_m\}$.

LEMMA 2.2. $Q_e = Q_D$.

Proof. Let $\mu \in Q_e$. Suppose that $\mu \notin Q_D$. Pick A and f as in Lemma 2.1. Then $\alpha_1(\nu(x, \cdot))$ and $\alpha_2(\nu(x, \cdot))$ must differ from each other for $x \in A$. Define μ_1 and μ_2 by

$$\mu_i(dx, dy) = \nu(dx) \alpha_i(\nu(x, dy)), \quad i = 1, 2.$$

By (2.2), $\mu = (\mu_1 + \mu_2)/2$. Clearly, $\mu_1 \neq \mu_2$. Thus $\mu \notin Q_e$, a contradiction. Thus $Q_e \subset Q_D$. Conversely, supposed that $\mu \in Q_D$ is of the form $\mu = (\mu_1 + \mu_2)/2$ for some $\mu_1 \neq \mu_2$ in Q . Then, for

$$\mu_i(dx, dy) = \nu(dx) v_i(x, dy), \quad i = 1, 2,$$

$v_1(x, \cdot)$ and $v_2(x, \cdot)$ must differ for x in a set of strictly positive ν -measure. For such x , however, $\nu(x, \cdot) = (v_1(x, \cdot) + v_2(x, \cdot))/2$ cannot be Dirac, a contradiction. Thus $Q_D \subset Q_e$. \square

We deduce some important consequences of this based on a small extension of Choquet's theorem, which follows.

LEMMA 2.3. *Let X be a Polish space, and let $G \subset P(X)$ be a closed convex set, with $G_e \subset G$ the set of its extreme points. Then every element of G is the barycenter of a probability measure on G_e .*

Proof. If G is compact, the proof is immediate from the classical Choquet's theorem [6]. As described in Remark 2.1, X is homeomorphic to a G_δ subset Y of $[0, 1]^\infty$. We identify X and Y and denote by \bar{X} the closure of X in $[0, 1]^\infty$. (Note that the corresponding identification between $P(X)$ and $P(Y)$ preserves extreme points of convex sets.) \bar{X} and therefore $P(\bar{X})$ are compact. View $P(X)$ as a subset of $P(\bar{X})$ by identifying each element μ of $P(X)$ with its unique extension $\bar{\mu}$ in $P(\bar{X})$, i.e., $\bar{\mu}$ restricts to μ on X and $\bar{\mu}(\bar{X} \setminus X) = 0$. Let \bar{G} be the closure of G in $P(\bar{X})$, and let \bar{G}_e be the set of its extreme points. By Choquet's theorem, each $\mu \in G$ is the barycenter of some $\eta \in P(\bar{G}_e)$. Fix μ and η . If an element e of G_e is not in \bar{G}_e , it must be a convex combination of two distinct elements of \bar{G} at least one of which must charge $\bar{X} \setminus X$. (Otherwise, both are in G , contradicting $e \in G_e$.) However, then $e(\bar{X} \setminus X) > 0$, contradicting $e \in G$. Thus $G_e \subset \bar{G}_e$. If $\eta(\bar{G}_e \setminus G_e) > 0$, $\mu(\bar{X} \setminus X) > 0$ because each $\mu' \in \bar{G}_e \setminus G_e$ satisfies $\mu'(\bar{X} \setminus X) > 0$. This is impossible, however. Hence $\eta(G_e) = 1$, proving the claim. \square

COROLLARY 2.1. *Each $\mu \in Q$ is the barycenter of a probability measure η supported on Q_D .*

This is immediate in view of the foregoing. Suppose now that S_1 above is of the form $S' \times S''$, where S' and S'' are Polish spaces. Also, let μ of (2.1) be of the form

$$(2.4) \quad \mu(dx, dy, dz) = \varphi(dy)\psi_1(y, dx)\psi_2(y, dz)$$

for $\varphi \in P(S'')$ and $\psi_1 : S'' \rightarrow P(S')$, $\psi_2 : S'' \rightarrow P(S_2)$ measurable.

COROLLARY 2.2. *The measure μ of (2.4) is the barycenter of a probability measure η on Q_D satisfying the additional property that η is supported on the subset of Q_D consisting of probability measures of the type*

$$(2.5) \quad \varphi(dy)\psi_1(y, dx)q(y, dz),$$

where $q(y, \cdot) \in P(S_2)$ is a Dirac measure for φ -almost surely y .

Proof. Applying Corollary 2.1 to the measure

$$\varphi(dy)\psi_2(y, dz)$$

on $S'' \times S_2$, we see that it is the barycenter of a probability measure on the set of probability measures of the type

$$(2.6) \quad \varphi(dy)q(y, dz),$$

where $q(y, \cdot)$ is Dirac for φ -almost surely y . Given the obvious one-to-one correspondence between measures (2.5) and (2.6), the claim follows. \square

3. White-noise representation. Let X and Y be random variables on some probability space (Ω, \mathcal{F}, P) , taking values in Polish spaces S_1 and S_2 , respectively.

LEMMA 3.1. *There exists a probability space $(\Omega', \mathcal{F}', P')$ with random variables X', Y' , and Z' defined on it, taking values in S_1 , S_2 , and $P(S_1 \times S_2)$, respectively, such that (a) (X', Y') agree in law with (X, Y) , (b) (Y', Z') are independent, and (c) $X' = f(Z', Y')$ for a measurable $f : P(S_1 \times S_2) \times S_2 \rightarrow S_1$.*

Proof. Let the μ of (2.1) be the law of (Y, X) , and let η be as in Corollary 2.1. Then η is supported on probability measures on $S_2 \times S_1$ of the type

$$\nu(dx)\delta_{g(x)}(dy),$$

where δ_z is the Dirac measure at z . Let $\Omega' = P(S_1 \times S_2) \times S_2 \times S_1$ with \mathcal{F}' = the product σ -field and P' = the probability measure on (Ω', \mathcal{F}') defined by

$$P'(d\rho, dx, dy) = \eta(d\rho)\rho(dx, dy) = \eta(d\rho)\nu(dx)\delta_{f(\rho, x)}(dy)$$

for a measurable $f: P(S_1 \times S_2) \times S_2 \rightarrow S_1$. Let (Z', Y', X') be the canonically realized random variables on this probability space. (That is, if $\omega = (\omega_1, \omega_2, \omega_3)$ is a typical element of Ω' , then $Z'(\omega) = \omega_1$, $Y'(\omega) = \omega_2$, and $X'(\omega) = \omega_3$.) The claim follows. \square

COROLLARY 3.1. *In Lemma 3.1, $P(S_1 \times S_2)$ may be replaced by any prescribed uncountable Polish space S (say, $[0, 1]$) without any loss of generality.*

Proof. We may suppose that S_1 and S_2 are not singletons. (If, say, $S_1 = \{a\}$, replace it by $S_1 = \{a, b\}$, $b \neq a$, with $P(X = b) = 0$.) Then $P(S_1 \times S_2)$ is uncountable. By the isomorphism theorem there exists a measurable isomorphism $g: P(S_1 \times S_2) \rightarrow S$. Letting $\tilde{Z} = g(Z')$ and $\tilde{f}(\cdot, \cdot) = f(g^{-1}(\cdot), \cdot): S \times S_2 \rightarrow S_1$, we have $X' = \tilde{f}(\tilde{Z}, Y')$. \square

COROLLARY 3.2. *In Lemma 3.1 we may take $S = [0, 1]$ and \tilde{Z} to be uniformly distributed.*

Proof. Let $S = [0, 1]$, and let \tilde{f} and \tilde{Z} be as in Corollary 3.1. Let $F: [0, 1] \rightarrow [0, 1]$ be the distribution function of \tilde{Z} . Then F is nondecreasing and right-continuous. Defining $q: [0, 1] \rightarrow [0, 1]$ by

$$q(x) = \min F^{-1}(\min(F([0, 1]) \cap [x, 1])),$$

we check that the law of \tilde{Z} is the image of the uniform measure on $[0, 1]$ under q . For a random variable T uniformly distributed on $[0, 1]$, let $x \mapsto r(x, du): [0, 1] \rightarrow P([0, 1])$ denote a version of the regular conditional law of T , given $q(T)$. Augment the probability space $(\Omega', \mathcal{F}', P')$ in Lemma 3.1 as follows: Replace Ω' by $\Omega' \times [0, 1]$, replace \mathcal{F}' by its product with the Borel σ -field of $[0, 1]$, and replace P' by the probability measure $P'(dw)r(\tilde{Z}(w), dw')$. Define a new $[0, 1]$ -valued random variable Z on this probability space by $Z((w, w')) = w'$. Then Z is uniformly distributed on $[0, 1]$ and $\tilde{Z} = q(Z)$. Thus $X' = f'(Z, Y')$, where $f'(\cdot, \cdot) = \tilde{f}(q(\cdot), \cdot)$. \square

COROLLARY 3.3. *In Lemma 3.1 we may take $X' = X$ and $Y' = Y$, and we may take Z' (or \tilde{Z} or Z) to be a random variable constructed on an augmentation of the probability space (Ω, \mathcal{F}, P) .*

Proof. To be specific, we prove the statement for Z' . Augment (Ω, \mathcal{F}, P) as follows: Replace Ω by $\Omega \times P(S_1 \times S_2)$, replace \mathcal{F} by its product with the Borel σ -field of $P(S_1 \times S_2)$, and replace P by a new probability measure defined as follows. Let $v: S_1 \times S_2 \rightarrow P(P(S_1 \times S_2))$ be a version of the regular conditional law of Z' , given X' and Y' . Replace P by the probability measure

$$P(dw)v((X(w), Y(w)), dz).$$

Define a $P(S_1 \times S_2)$ -valued random variable \tilde{Z} on this new probability space by $\tilde{Z}((w, z)) = z$, where (w, z) is a typical sample point of $\Omega \times P(S_1 \times S_2)$. By construction, the laws of (X', Y', Z') and (X, Y, \tilde{Z}) agree. \square

The implications for stochastic realization theory follow. Let U be an uncountable Polish space.

THEOREM 3.1. *Let $\{X_n, n = 0, 1, 2, \dots\}$ be a random sequence taking values in a Polish space S and defined on a probability space (Ω, \mathcal{F}, P) . Then, on another probability space $(\Omega', \mathcal{F}', P')$, we can construct S -valued random variables $\{X'_n, n = 0, 1, 2, \dots\}$ and U -valued independent and identically distributed random variables $\{W_n\}$ such that $\{X_n\}$ and $\{X'_n\}$ agree in law, W_n is independent of $(X'_{n-1}, W_{n-1}, X'_{n-2}, W_{n-2}, \dots, X'_0, W_0)$ for each n and*

$$(3.1) \quad X'_n = f_n(W_n, X'_{n-1}, X'_{n-2}, \dots, X'_0)$$

for measurable $f_n: U \times S^n \rightarrow S$, $n = 1, 2, \dots$. Furthermore, if $\{X_n\}$ is Markov, (3.1) may be replaced by

$$(3.2) \quad X'_n = f_n(W_n, X'_{n-1})$$

for measurable $f_n: U \times S \rightarrow S$, $n = 1, 2, \dots$.

Proof. We first prove (3.1) for independent but not necessarily identically distributed $\{W_n\}$ with W_n independent of $(X'_{n-1}, X'_{n-2}, \dots)$ for each n . For $n = 1$ the claim follows from Lemma 3.1. Suppose that it is true for $n \leq m$. Let μ denote the law of $(X_0, \dots, X_m, W_1, \dots, W_m)$, written as

$$\begin{aligned} \mu(dx_0, \dots, dx_m, dw_1, \dots, dw_m) \\ = \nu(dx_0, \dots, dx_m)q((x_0, \dots, x_m), (dw_1, \dots, dw_m)), \end{aligned}$$

where ν is the law of (X_0, \dots, X_m) and $(x_0, \dots, x_m) \rightarrow q((x_0, \dots, x_m), (dw_1, \dots, dw_m))$ is a version of the regular conditional law of (W_1, \dots, W_m) , given (X_0, \dots, X_m) . Consider the probability measure

$$\begin{aligned} \bar{\mu}(dx_0, \dots, dx_{m+1}, dw_1, \dots, dw_m) \\ = \nu(dx_0, \dots, dx_m)q((x_0, \dots, x_m), (dw_1, \dots, dw_m))r((x_0, \dots, x_m), dx_{m+1}), \end{aligned}$$

where $(x_0, \dots, x_m) \rightarrow r((x_0, \dots, x_m), dx_{m+1})$ is the regular conditional law of X_{m+1} , given (X_0, \dots, X_m) . By Corollary 2.2 this is the barycenter of a measure η on $P(S^{m+2} \times U^m)$ concentrated on probability measures of the type

$$\nu(dx_0, \dots, dx_m)q((x_0, \dots, x_m), (dw_1, \dots, dw_m))\delta_{f(x_0, \dots, x_m)}(dx_{m+1}).$$

Consider $P(S^{m+2} \times U^m) \times S^{m+1} \times U^m \times S$ with the probability measure

$$\begin{aligned} \eta(d\rho)\rho(dx_0, \dots, dx_m, dw_1, \dots, dw_m, dx_{m+1}) \\ = \eta(d\rho)\mu(dx_0, \dots, dx_m, dw_1, \dots, dw_m)\delta_{\varphi(\rho, x_m, \dots, x_0)}(dx_{m+1}) \end{aligned}$$

for a measurable $\varphi: P(S^{m+2} \times U^m) \times S^{m+1} \rightarrow S$. If $(\bar{W}_{m+1}, X'_0, \dots, X'_m, W'_1, \dots, W'_m, X'_{m+1})$ denote the canonically realized random variables on this space, it follows by construction that

$$X'_{n+1} = \varphi(\bar{W}_{n+1}, X'_n, \dots, X'_n)$$

and that \bar{W}_{n+1} is independent of $X'_0, \dots, X'_n, W'_1, \dots, W'_n$. The desired claim now follows by induction. By adapting the argument of Corollary 3.1, we may replace $P(S^{n+2} \times U^n)$, $n = 0, 1, 2, \dots$, by $[0, 1]$ and then use the construction of Corollary 3.2 to replace $\{\bar{W}_n\}$ by independent and identically distributed random variables uniformly distributed on $[0, 1]$. In turn, the isomorphism theorem allows us to replace these by independent and identically distributed random variables taking values in any prescribed uncountable Polish space as in the proof of Corollary 3.1. Equation (3.2) is also proved by analogous arguments in view of Corollary 2.2. \square

THEOREM 3.2. Let $\{X_n, n = 0, \pm 1, \pm 2, \dots\}$ be a sequence of S -valued random variables on a probability space (Ω, \mathcal{F}, P) . Then we can construct on some probability space $(\Omega', \mathcal{F}', P')$ a sequence of S -valued random variables $\{X'_n, n = 0, \pm 1, \pm 2, \dots\}$ and U -valued independent and identically distributed random variables $\{W_n, n = 0, \pm 1, \pm 2, \dots\}$ such that $\{X_n\}$ and $\{X'_n\}$ agree in law, W_n is independent of $(X_{n-1}, W_{n-1}), (X_{n-2}, W_{n-2}), \dots$ for each n , and

$$X_n = f_n(W_n, X_{n-1}, X_{n-2}, \dots)$$

for measurable $f_n: U \times S^\infty \rightarrow S$, $n = 0, \pm 1, \pm 2, \dots$. Furthermore, if $\{X_n\}$ is a stationary process, $\{f_n\}$ may be taken to be independent of n .

Proof. Proceeding as in Theorem 3.1, we can construct a sequence of processes $(X_n^m, W_n^m, n = 0, \pm 1, \pm 2, \dots)$, $m = 1, 2, \dots$, such that $\{X_n^m\}$ and $\{X_n\}$ agree in law, $W_n^m, n = -m, -m + 1, -m + 2, \dots$ is an independent and identically distributed sequence uniformly distributed on $[0, 1]$ and satisfying the following: W_{-m+i}^m is independent of $(X_{-m+i-1}^m, W_{-m+i-1}^m), (X_{-m+i-2}^m, W_{-m+i-2}^m), \dots$, for each $i \geq 0$, $W_n^m =$ a prescribed element u of $[0, 1]$ for $n < -m$, and for each $n \geq -m$

$$X_n^m = f_n(W_n^m, X_{n-1}^m, X_{n-2}^m, \dots).$$

(We simply mimic the steps of Theorem 3.1 with $(X_0^m, X_{-1}^m, X_{-2}^m, \dots)$ in place of X_0^m . That f_n can be taken to be independent of m is a consequence of the fact that this choice depends purely on the joint law of $(X_n^m, X_{n-1}^m, X_{n-2}^m, \dots)$, which is the same as that of (X_n, X_{n-1}, \dots) and is thus independent of m .) As $m \rightarrow \infty$, this sequence of processes is seen to converge in law to a process $(X_n', W_n, n = 0, \pm 1, \pm 2, \dots)$, which satisfies the requirements of the first claim. The second claim then follows from the observation that the possible choices of f_n are dictated by the joint law of $X_n, X_{n-1}, X_{n-2}, \dots$, which is independent of n for a stationary process. Finally, the isomorphism theorem allows us, as usual, to replace $\{W_n\}$ above by independent and identically distributed random variables taking values in any prescribed uncountable Polish space U . \square

COROLLARY 3.4. Let $(X_n, n = 0, \pm 1, \pm 2, \dots)$ be a stationary Markov process on a probability space (Ω, \mathcal{F}, P) . Then there exist on some probability space $(\Omega', \mathcal{F}', P')$ processes $(X_n', W_n, n = 0, \pm 1, \pm 2, \dots)$ such that $\{X_n'\}$ and $\{X_n\}$ agree in law, $\{W_n\}$ are independent and identically distributed U -valued random variables with W_n independent of $(X_{n-1}', W_{n-1}'), (X_{n-2}', W_{n-2}', \dots)$ for each n , and $X_n = f(W_n, X_{n-1})$ for each n and some measurable $f: U \times S \rightarrow S$.

This is obvious from the foregoing. Recall the definitions of a weak realization (ii) and weak white-noise representation (iv).

THEOREM 3.3. Given a weak stochastic realization $(\{X_n'\}, \{Y_n'\})$ of a random sequence $\{Y_n\}$ of S -valued random variables, a weak white-noise representation as in (1.1)–(1.2) exists. At least one weak realization exists. Moreover, if $\{Y_n, n = 0, \pm 1, \pm 2, \dots\}$ is stationary, one exists in which $\{X_n\}$ is also stationary.

Proof. The first claim is proved by a straightforward adaptation of the proofs of Theorems 3.1 and 3.2. The second claim follows from the trivial observation that $X_n = [Y_n, Y_{n-1}, Y_{n-2}, \dots]$ (with $Y_{-m}, m \leq 1$, equal to a fixed element of S when the original $\{Y_n\}$ are defined for only $n = 0, 1, 2, \dots$) is an S^∞ -valued Markov process and $Y_n = g(X_n)$, where $g: S^\infty \rightarrow S$ is the projection onto the first factor space. The last claim also follows similarly. \square

Remark 3.1. As in Corollary 3.3, we may construct the processes $\{X_n'\}$ and $\{W_n'\}$ on an augmentation of the original probability space (Ω, \mathcal{F}, P) . This augmentation in general cannot be avoided, as the following example shows. (Thus the strong white-noise representation is not always feasible.) Let X and Y be random variables taking values in $\{a, b\}$ and $\{c, d\}$, respectively, with

$$P(Y = c) = P(Y = d) = \frac{1}{2},$$

$$P(X = a/Y = c) = \frac{1}{2} = P(X = b/Y = c),$$

$$P(X = a/Y = d) = \frac{1}{3} = 1 - P(X = b/Y = d).$$

The law of (X, Y) is then the barycenter of a probability measure supported on the four probability measures $\{\mu_1, \mu_2, \mu_3, \mu_4\}$ on $\{a, b\} \times \{c, d\}$ described as follows:

$$\mu_i(Y = c) = \mu_i(Y = d) = \frac{1}{2} \text{ for } 1 \leq i \leq 4 \text{ and}$$

$$\mu_1(X = a/Y = c) = 1 = \mu_1(X = b/Y = d),$$

$$\mu_2(X = b/Y = c) = 1 = \mu_2(X = b/Y = d),$$

$$\mu_3(X = a/Y = c) = 1 = \mu_3(X = a/Y = d),$$

$$\mu_4(X = b/Y = c) = 1 = \mu_4(X = a/Y = d).$$

In fact, $\eta(\{\mu_1\}) = \eta(\{\mu_2\}) = \frac{1}{3}$, and $\eta(\{\mu_3\}) = \eta(\{\mu_4\}) = \frac{1}{6}$.

As above, we can construct on a possibly augmented probability space a random variable W taking values in $\{\mu_1, \mu_2, \mu_3, \mu_4\}$ with law η such that $X = f(W, Y)$ for a suitably defined f . A simple calculation using the Bayes rule shows that $P(W = \mu_i/X = a, Y = c) > 0$ for $i = 1, 3$. Thus W cannot be expressed as a function of (X, Y) . Suppose that (X, Y) were canonically realized on their canonical space $\{a, b\} \times \{c, d\}$. Then any random variable defined on this space must be a function of (X, Y) , and therefore W cannot be realized on this space unless we allow for its augmentation.

4. Connections with the ergodic decomposition of stationary processes. This section proves another conjecture of Willems and Van Schuppen [10] concerning white-noise representations of stationary sequences (see [7] for some related work). Let $X_n, n = 0, \pm 1, \pm 2, \dots$ be a stationary sequence of S -valued random variables. As in Theorem 3.2, we have

$$X_n = f(W_n, X_{n-1}, X_{n-2}, \dots), \quad n = 0, \pm 1, \pm 2, \dots,$$

where $\{W_n\}$ are independent and identically distributed U -valued random variables such that W_n is independent of $(X_{n-1}, W_{n-1}), (X_{n-2}, W_{n-2}), \dots$. Thus

$$\begin{aligned} X_n &= f(W_n, f(W_{n-1}, X_{n-1}, X_{n-2}, \dots), f(W_{n-2}, X_{n-2}, X_{n-3}, \dots), \dots) \\ &= f(W_n, f(W_{n-1}, f(W_{n-2}, X_{n-2}, \dots)), f(W_{n-2}, f(W_{n-3}, X_{n-3}, \dots)), \dots). \end{aligned}$$

Iterating, we may hope to get

$$(4.1) \quad X_n = F(W_n, W_{n-1}, W_{n-2}, \dots), \quad n = 0, \pm 1, \pm 2, \dots,$$

for a measurable $F: U^\infty \rightarrow S$. As observed in [10], this may not always be possible. We prove below that such a representation holds if and only if $\{X_n\}$ is an ergodic sequence, as conjectured in [10]. For simplicity, we set $U = [0, 1]$ and assume $\{W_n\}$ to be uniformly distributed on $[0, 1]$. This causes no loss of generality, as already observed. We write $X = [\dots, X_{n-1}, X_n, X_{n+1}, \dots]$ and $W = [\dots, W_{n-1}, W_n, W_{n+1}, \dots]$. Also, let θ denote the shift operator on S^∞ or U^∞ , as the case may be, which maps $(\dots, x_{n-1}, x_n, \dots)$ into $(\dots, x_n, x_{n+1}, \dots)$.

In the setup of §2, let $S_1 = U^\infty$ and $S_2 = S^\infty$, and let Q be the set of $\mu \in P(S_1 \times S_2)$ satisfying the following: μ disintegrates as in (2.1) with both ν and $x \rightarrow v(x, \cdot)$ invariant under θ (that is, $\nu(A) = \nu(\theta^{-1}(A))$ and $v(x, B) = v(\theta x, \theta^{-1}(B))$ for all Borel $A \subset U^\infty$ and $B \subset S^\infty$ and for all $x \in U^\infty$). Argue as in §2 to show that the extreme points of Q are precisely those μ for which $x \rightarrow v(x, \cdot)$ is Dirac for ν -almost surely x (and, of course, θ -invariant). Thus each $\bar{\mu} \in Q$ is the barycenter of a probability measure on the set of such measures. Argue as for Lemma 3.1 and Corollary 3.3 to conclude that

$$(4.2) \quad X = g(Z, W)$$

for a random variable Z independent of W and uniformly distributed on $[0, 1]$, defined possibly on an augmentation of the original probability space, and a measurable $g : [0, 1] \times U^\infty \rightarrow S^\infty$, which, in view of the above discussion, satisfies

$$\theta g(\cdot, \cdot) = g(\cdot, \theta(\cdot)).$$

Hence (4.2) is equivalent to the infinitely many equations

$$(4.3) \quad X_n = f(Z, \theta^n(W)), \quad n = 0, \pm 1, \pm 2, \dots,$$

where $f = p \circ g$, where $p : S^\infty \rightarrow S$ is the projection $(\dots, x_{-1}, x_0, x_1, \dots) \rightarrow x_0$. For $z \in [0, 1]$, define

$$X^z = [\dots, X_{n-1}^z, X_n^z, X_{n+1}^z, \dots]$$

by

$$X_n^z = f(z, \theta^n(W)), \quad n = 0, \pm 1, \pm 2, \dots$$

Then $\{X_n^z, n = 0, \pm 1, \pm 2, \dots\}$ are stationary processes for $z \in [0, 1]$ with laws denoted by, say, $L_z \in P(S^\infty)$. By (4.3), $L =$ (the law of X) is the barycenter of a probability measure η on $L_z, z \in [0, 1]$. Suppose that L_z is not a constant element of $P(S^\infty)$ for almost every $z \in [0, 1]$. Then there exist $m \geq 1$ and $f \in C_b(S^m)$ such that

$$h(z) = \int f(x_1, \dots, x_m) dL_z(\dots, x_{-1}, x_0, x_1, \dots), \quad z \in [0, 1]$$

is not a constant Lebesgue almost everywhere. It is clear that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^{n-1} f(X_{i+1}, X_{i+2}, \dots, X_{i+m}) = h(Z)$$

almost surely. The left-hand side is a random variable measurable with respect to the shift-invariant σ -field of X (consisting of all events of the type $\{X \in A\}$, A Borel in S^∞ , that satisfy $\{X \in A\} = \{\theta^n X \in A\}$ for all n). Since it is not a constant almost surely, this σ -field is not trivial, and therefore X cannot be ergodic. Hence, if X is ergodic, L_z must be a constant element of $P(S^\infty)$ for almost every z , and thus $X' = g(z, W)$ is a replica in law of X for almost every z in $[0, 1]$. On the other hand, if L_z is a constant element of $P(S^\infty)$ for almost every z in $[0, 1]$, then $X' = g(z, W)$ is a replica in law of X for almost every z in $[0, 1]$. The shift-invariant σ -field of X' (consisting of all events of the type $\{X' \in A\}$, A Borel on S^∞ , that satisfy $\{X' \in A\} = \{\theta^n X' \in A\}$ for all n) is contained in the shift-invariant σ -field of W . However, W is an ergodic process, and therefore its shift-invariant σ -field is trivial. Thus the shift-invariant σ -field of X' is trivial, and X' and therefore X are ergodic. We have proved that $X_n = f(\theta^n W)$ for all n and some measurable $f : U^\infty \rightarrow S$ if and only if X is ergodic. However, for each n , $(W_{n+1}, W_{n+2}, \dots)$ is independent of X_n, W_n, W_{n-1}, \dots . Thus the law of $[(W_{n+1}, W_{n+2}, \dots), (W_n, W_{n-1}, \dots), (X_n)]$ is of the form

$$\varphi_1(dW_{n+1}, dW_{n+2}, \dots) \varphi_2(dW_n, dW_{n-1}, \dots) q((W_n, W_{n-1}, \dots), dx_n)$$

for $\varphi_1, \varphi_2 \in P(U^\infty)$, $q : U^\infty \rightarrow P(S)$. If X_n is a function of $\theta^n W$, $q((W_n, W_{n-1}, \dots), dx_n)$ must be a Dirac measure at, say, $f'(W_n, W_{n-1}, \dots)$. However, then $X_n = f'(W_n, W_{n-1}, \dots)$. This completes the proof of our desired result.

THEOREM 4.1. $\{X_n\}$ has a representation (on a possibly augmented probability space) of the type (4.1) if and only if it is ergodic.

In conclusion, we remark on some important issues not touched on in this paper. The first, of course, is the minimal realization problem mentioned in the Introduction. It is not clear whether an explicit white-noise representation will contribute to its resolution. In the context of the white-noise representation in (1.1) and (1.2) (for a given weak realization), we can pose some interesting minimality questions. Consider the stationary Markov case for simplicity. The representation then is

$$X_n = f(W_n, X_{n-1}).$$

Both f and the distribution F of W_n are not unique, as already seen. Can we find the “best” pair (f, F) with respect to a suitable complexity measure? Obviously, “best” would mean “minimum complexity.” Possibly there would be a tradeoff between the (as yet undefined) complexities of f and F , which we can hope to quantify.

Another relevant issue not touched on here is whether we can give conditions on the law of $\{X_n\}$ that ensure a suitable level of regularity (continuity, Lipschitz property, and so forth) of the functions $\{f_n\}$ and $\{g_n\}$ featured in (1.1) and (1.2).

Note added in proof. After the editorial process for this paper was over, the author came across a paper by Brown [11] whose Lemma 1, p. 178 gives an easy proof of Theorem 3.1, above, for $[0, 1]$ -valued random variables. The general case may then be deduced by invoking the isomorphism theorem. The proof given in the present paper, however, is somewhat more direct in the case of general Polish spaces and may be of some value in pursuing the open issues listed in the concluding paragraphs of the present paper.

REFERENCES

- [1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [2] V. BORKAR, *On extremal solutions to stochastic control problems*, Appl. Math. Optim., 24 (1991), pp. 317–330.
- [3] L. DUBINS AND D. FREEDMAN, *Measurable sets of measures*, Pacific J. Math., 14 (1964), pp. 1211–1222.
- [4] A. LINDQUIST AND G. PICCI, *On the stochastic realization problem*, SIAM J. Control Optim., 17 (1979), pp. 365–389.
- [5] K. R. PARTHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.
- [6] R. PHELPS, *Lectures on Choquet's Theorem*, Van Nostrand, New York, 1966.
- [7] M. ROSENBLATT, *Stationary processes as shifts of functions of independent random variables*, J. Math. Mech., 8 (1959), pp. 665–681.
- [8] L. SCHWARTZ, *Disintegration of measures*, Tata Institute of Fundamental Research, Bombay, 1976.
- [9] J. VAN SCHUPPEN, *Stochastic realization problems*, in Three Decades of Mathematical System Theory, H. Nijmeijer and J. M. Schumacher, eds., Lecture Notes in Control and Information Sciences 135, Springer-Verlag, Berlin, Heidelberg, New York, 1989, pp. 480–523.
- [10] J. WILLEMS AND J. VAN SCHUPPEN, *Stochastic systems and the problem of state space realization*, NATO ASI-AMS Seminar on Algebraic and Geometric Methods in Linear System Theory, Harvard University, Cambridge, MA, June 1979.
- [11] T. C. BROWN, *Poisson approximations and exchangeable random variables*, in Exchangeability in Probability and Statistics, G. Koch and F. Spizzichino, eds., North-Holland, Amsterdam, 1982, pp. 177–183.

THE MINIMAL TIME FUNCTION IN INFINITE DIMENSIONS*

OVIDIU CÂRJĂ†

Abstract. The goal of this paper is to study regularity properties of the minimal time function $T(\cdot)$ associated with linear control systems in infinite dimensions. This paper proves that various regularity properties of $T(\cdot)$ hold on the whole reachable set provided they hold around the target. Some methods to obtain estimates on $T(\cdot)$ around the target are given and applications to both distributed and boundary control problems are presented.

Key words. infinite-dimensional control problem, minimal time function

AMS subject classifications. 49K40, 49N05

1. Introduction. Consider a control system represented by

$$(1.1) \quad \begin{aligned} y(t, x, u) &= S(t)x + H(t)u, \quad t > 0 \\ y(0, x, u) &= x, \end{aligned}$$

where y is the state, t the time, and u the control. Here, $S(t)$, $t \geq 0$, is a C_0 -semigroup on a Banach space X and $H(t)$, $t > 0$, is a family of bounded linear operators, $H(t) : L^p(0, t; U) \rightarrow X$, such that the following condition is satisfied:

$$(1.2) \quad \begin{aligned} H(t_1 + t_2)u &= S(t_2)H(t_1)u + H(t_2)J_{t_1}u, \\ \text{for all } t_1, t_2 > 0, \quad u &\in L^p(0, t_1 + t_2, U), \end{aligned}$$

where U is a Banach space, $1 \leq p \leq \infty$, and $\{J_s, s \in \mathbb{R}\}$ is a family of translation operators, $J_s : L^p_{\text{loc}}(\mathbb{R}^+; U) \rightarrow L^p_{\text{loc}}(\mathbb{R}^+; U)$,

$$(J_s u)(t) = u(t + s) \quad \text{for } s \geq 0 \quad \text{and}$$

$$(J_s u)(t) = \begin{cases} 0 & \text{if } 0 \leq t \leq |s| \\ u(t + s) & \text{if } t > |s|, \end{cases} \quad \text{for } s < 0.$$

Of course, $H(t_1)u$ in (1.2) means $H(t_1)P_{t_1}u$, where $P_{t_1}u$ is the projection of $L^p(0, t_1 + t_2, U)$ onto $L^p(0, t_1; U)$, i.e.,

$$(P_{t_1}u)(s) = \begin{cases} u(s) & \text{if } 0 \leq s \leq t_1 \\ 0 & \text{if } s \geq t_1 \end{cases}; \text{ similarly for } H(t_2)J_{t_1}u.$$

A nice representation theorem for a family $H(t)$ that satisfies (1.2) is given in [25], but we do not use this fact here.

There is a large class of control systems that can be represented as in (1.1). We give some examples at the end of this section.

Let us fix a target $x_1 \in X$ and an admissible set of controls $U_{\text{ad}} \in L^p_{\text{loc}}(\mathbb{R}^+; U)$. For $t \geq 0$, denote by $V(t)$ the set of states controllable to x_1 within time t by admissible controls; that is

$$V(t) = \{x \in X; y(t, x, u) = x_1 \text{ for some } u \in U_{\text{ad}}\}.$$

* Received by the editors February 25, 1991; accepted for publication (in revised form) March 3, 1992.

† Department of Mathematics, University of Iași, 6600 Iași, Romania.

The set of states controllable to x_1 in free time by admissible controls is defined by

$$V = \bigcup_{t \geq 0} V(t).$$

Finally, define the minimal time function (the Bellman function) $T: X \rightarrow [0, \infty]$

$$(1.3) \quad \begin{aligned} T(x) &= \inf \{t; x \in V(t)\} \quad \text{for } x \in V \\ T(x) &= \infty \quad \text{for } x \notin V. \end{aligned}$$

The goal of this paper is to study some continuity properties of the minimal time function defined above. There is an extensive literature on the minimal time function associated with finite-dimensional control systems (see, e.g., [3] and the bibliography given there). The motivation for this interest comes from the fact that this function is used in the construction of feedback controls (see [13], [17, p. 146]) and that its properties provide a basis in the study of the dynamic programming equation (see [17], [3]). Moreover, the local growth of the minimal time function is closely related to the sensitivity of the optimal time with respect to perturbations in the initial state. However, in the infinite-dimensional case there are few papers that concern this subject (see [1], [5], [6]).

Very recently, Barbu [2] studied the minimal time function associated to some nonlinear infinite-dimensional control systems in connection with the theory of viscosity solution to the Bellman equation.

This paper is organized as follows. In § 2 we prove that various regularity properties of the minimal time function, such as local boundedness and uniform and Lipschitz continuity hold on the whole reachable set provided they hold around the target. In § 3, we relate the continuity of the minimal time function to the null controllability of (1.1) by unconstrained controls in arbitrarily short time and provide practical techniques to achieve estimates on $T(\cdot)$ around the target, in the case where $x_1 = 0$. Finally, in § 4, we apply the general theory to concrete examples.

This section concludes with some examples of control systems of type (1.1), which satisfy (1.2).

Example 1.1. Let $B \in L(U, X)$ and $H(t) = \int_0^t S(t-s)Bu(s) ds$, where $S(t)$ is a C_0 -semigroup on the Banach space X . It is easy to verify condition (1.2) for every $p \in [1, \infty]$. In this case, (1.1) gives the mild solution to the equation $y' = Ay + Bu$, $y(0) = x$.

Example 1.2. The heat equation with boundary control can also be considered as a control system of type (1.1) (see, e.g., [20]).

Example 1.3. Consider the wave equation defined on a smooth bounded domain $\Omega \subset \mathbb{R}^n$,

$$(1.4) \quad \begin{aligned} (a) \quad & y'' - \Delta y = 0 \quad \text{in } \Omega \times (0, t) \\ (b) \quad & y = u \quad \text{in } \partial\Omega \times (0, t) = \Sigma \\ (c) \quad & y(0) = y_0; \quad y'(0) = y_1 \quad \text{in } \Omega. \end{aligned}$$

Defining for $t \geq 0$, $S(t): L^2(\Omega) \times H^{-1}(\Omega) \rightarrow L^2(\Omega) \times H^{-1}(\Omega)$, $S(t)(y_0, y_1) = (y(t), y'(t))$, where $y(\cdot)$ is the solution to (1.4a), (1.4c) with $y = 0$ on Σ , $S(t)$ is a C_0 -semigroup on $L^2(\Omega) \times H^{-1}(\Omega)$ (see [24]). On the other hand, $H(t): L^2(\Sigma) \rightarrow L^2(\Omega) \times H^{-1}(\Omega)$, $H(t)u = (y(t), y'(t))$, where $y(\cdot)$ is the solution of (1.4a), (1.4b) with the initial condition $y(0) = 0$, $y'(0) = 0$, is bounded and, together with $S(t)$, verifies (1.2). Clearly, the solution of (1.4) may be written as (1.1) and condition (1.2) is verified with $p = 2$ and $U = L^2(\partial\Omega)$.

2. Qualitative properties of $T(\cdot)$. Throughout this section, we refer to a control system of type (1.1) with condition (1.2) and the associated minimal time function given by (1.3) corresponding to the admissible set of controls

$$(2.1) \quad U_{\text{ad}} = \bigcup_{t \geq 0} \{u \in L^p(0, t; U); \|u\|_p \leq \rho\},$$

where ρ is a given constant and $1 \leq p \leq \infty$. We begin with the following lemma.

LEMMA 2.1 (*Bellman optimality principle*). *For every $x \in V$ and $t \in [0, T(x)]$ we have*

$$T(x) = \inf \{t + T(y(t, x, u)); u \in U_{\text{ad}}\}.$$

Proof. First, we prove that

$$(2.2) \quad T(x) \leq t + T(y(t, x, u)) \quad \text{for every } u \in U_{\text{ad}}.$$

If $y(t, x, u) \notin V$, the inequality is clear. If $y(t, x, u) \in V(s)$ for some $s > 0$, then there exists $v \in U_{\text{ad}}$ such that $S(s)y(t, x, u) + H(s)v = x_1$. That is,

$$S(s)S(t)x + S(s)H(t)u + H(s)v = x_1.$$

By (1.2), the above relation may be written as

$$S(t + s)x + H(t + s)w = x_1,$$

where $w(\tau) = u(\tau)$, for $\tau \in [0, t]$ and $w(\tau) = v(\tau - t)$ for $\tau \in [t, t + s]$. Thus, $x \in V(t + s)$ and $T(x) \leq t + s$. This inequality holds for every s for which $y(t, x, u) \in V(s)$. Hence, (2.2) follows. To continue, let $\varepsilon > 0$. We have to prove that there exists $u \in U_{\text{ad}}$ such that $t + T(y(t, x, u)) \leq T(x) + \varepsilon$. Indeed, by the definition of $T(x)$, there exists $s > 0$ such that $s < T(x) + \varepsilon$ and $x \in V(s)$. Hence, there exists $v \in U_{\text{ad}}$ such that $S(s)x + H(s)v = x_1$. Since $t \leq T(x)$ we have that $t \leq s$. For $t = s$, all is clear. If $t < s$, by (1.2) we have

$$S(s - t)S(t)x + S(s - t)H(t)v + H(s - t)J_1v = x_1.$$

This shows that

$$S(t)x + H(t)v \in V(s - t),$$

hence, $T(y(t, x, v)) \leq s - t$. That is,

$$t + T(y(t, x, v)) \leq s < T(x) + \varepsilon.$$

The desired u is $P_t v$. The proof is complete.

Theorem 2.1, which follows, is the main result of this section. It relates the continuity of $T(\cdot)$ in x_1 to the same property on the whole set V . Moreover, a precise estimate is performed. The notation $\text{sign } \tau$ designates 1 for $\tau > 0$, -1 for $\tau < 0$, and 0 for $\tau = 0$.

THEOREM 2.1. *Assume $x_1 \in \text{int } V$. Then*

- (i) V is open.
- (ii) If, in addition, the minimal time function is bounded on a neighborhood of x_1 , then it is locally bounded on V .
- (iii) If, in addition, the minimal time function is continuous in x_1 , then it is locally uniformly continuous on V . More precisely, the following estimate holds:

$$(2.3) \quad |T(x) - T(y)| \leq T(h(x, y)) \quad \text{for } x, y \in V,$$

where $h(x, y) = x_1 + \text{sign}(T(x) - T(y))S(\min\{T(x), T(y)\})(x - y)$.

- (iv) If, in addition, the minimal time function is locally Lipschitz continuous in x_1 , that is, $T(x) \leq M\|x - x_1\|$, for some constant $M > 0$ and for every x in a neighborhood

of x_1 , then it is locally Lipschitz continuous on V , i.e., for every $x \in V$ there exist a neighborhood \mathcal{U} of x and a constant $M > 0$ such that

$$|T(y) - T(z)| \leq M\|y - z\| \quad \text{for every } y, z \in \mathcal{U}.$$

Proof. Suppose $x_1 \in \text{int } V$ and let us prove that V is open. To this end, let $z \in V$. More exactly, say $z \in V(t)$, and let us prove that $z \in \text{int } V$. Since V is a neighborhood of x_1 , taking into account that the operator $S(t)$ is bounded, we deduce the existence of a neighborhood \mathcal{U} of z such that $S(t)(x - z) + x_1 \in V$ for any $x \in \mathcal{U}$. Now, since $z \in V(t)$, there exists $v \in U_{\text{ad}}$ such that $y(t, z, v) = x_1$. Clearly, $S(t)(x - z) + x_1 = y(t, x, v)$. By Lemma 2.1, we obtain

$$(2.4) \quad T(x) \leq t + T(S(t)(x - z) + x_1), \quad x \in \mathcal{U},$$

which implies that $x \in V$. Hence, $\mathcal{U} \subset V$ and part (i) is proved. Part (ii) follows from (2.4). Let us now prove part (iii). Suppose that $T(\cdot)$ is continuous in x_1 and let $z \in V$. Consider \mathcal{U} and t as above. By (2.4), since $T(x_1) = 0$, we may infer that $\limsup_{x \rightarrow z} T(x) \leq t$. Since the above inequality is true for all $t > 0$ for which $z \in V(t)$, we obtain

$$\limsup_{x \rightarrow z} T(x) \leq T(z).$$

On the other hand, since V is open, there exists $\tau > T(z)$ such that $S(t)(x - z) + x_1 \in V$ for every $t \in (T(z), \tau)$. Consider $t \in (T(z), \tau)$ such that $z \in V(t)$. By (2.4) we obtain

$$T(x) - T(z) \leq \limsup_{t \downarrow T(z)} T(S(t)(x - z) + x_1) \leq T(S(T(z))(x - z) + x_1).$$

Interchanging the role of x and z , we get (2.3), as claimed. Combining (2.3) with the local boundedness of $T(\cdot)$ on V and with its continuity in x_1 we obtain that the minimal time function is locally uniformly continuous on V , and the proof of (iii) is completed. Finally, part (iv) follows easily from (iii).

We can now state a better result when $S(t)$ is a contraction semigroup.

COROLLARY 2.1. *Assume $x_1 \in \text{int } V$. Assume further that $S(t)$ is a contraction semigroup. Then, if the minimal time function is continuous in x_1 , then it is uniformly continuous on V . If, in addition, the minimal time function is locally Lipschitz continuous in x_1 , then it is Lipschitz continuous on V .*

Proof. The first part follows from Theorem 2.1. If $T(\cdot)$ is locally Lipschitz continuous in x_1 , then $L(x_1) < \infty$, where

$$L(x) = \limsup_{y \rightarrow x} \frac{|T(y) - T(x)|}{\|y - x\|}.$$

Let $x \in V$. From Theorem 2.1, $T(\cdot)$ is continuous on V , which is open. Hence, for y near x we have $y \in V$ and $h(x, y) \in V$. For such y , taking into account (2.3) we may infer that

$$\frac{|T(y) - T(x)|}{\|y - x\|} \leq \frac{T(h(x, y))}{\|y - x\|}.$$

Since $T(x_1) = 0$ and

$$\limsup_{y \rightarrow x} \frac{\|h(x, y) - x_1\|}{\|y - x\|} \leq 1$$

we deduce

$$L(x) \leq L(x_1) \quad \text{for every } x \in V.$$

We combine now this fact with the obvious inequality

$$|T(y) - T(x)| \leq \sup \{L(h); h \in [x, y]\} \|x - y\|,$$

for $x, y \in V$, to obtain that

$$|T(y) - T(x)| \leq L(x_1) \|y - x\|, \quad x, y \in V.$$

Here $[x, y]$ designates the entire segment from x to y . The proof is complete.

Remark 2.1. The proof of Theorem 2.1 uses primarily the Bellman optimality principle (Lemma 2.1). Because of this fact, similar results can be proved also for nonlinear control systems. Roughly speaking, we have the following principle: If a suitable continuity property of the solution with respect to the initial data holds true and the Bellman optimality principle is verified, then a continuity property of the minimal time function around the target implies a similar property on the whole reachable set (see [5]).

Remark 2.2. If the semigroup $S(t)$ is compact, then the continuity of $T(\cdot)$ in x_1 implies the weak continuity on V . See [1] and [2] for such results.

3. Estimates around the target. This section gives some general methods to obtain estimates for the minimal time function around the target. We consider here only the case $x_1 = 0$.

Let us recall first some results on constrained controllability for abstract linear control systems. In fact, these are results from operator theory.

LEMMA 3.1 (see [7]). *Let X, Y, Z be Banach spaces. Let $C : D(C) \subset X \rightarrow Y$ be linear with dense domain, and let $F \in L(X, Z)$. Then, the following conditions are equivalent:*

- (a) $F^*(Z^*) \subseteq C^*(Y^*)$;
- (b) *there exists $k_1 > 0$ such that $\|Fx\| \leq k_1 \|Cx\|, x \in D(C)$;*
- (c) *there exists $k_2 > 0$ such that*

$$\{F^*z^*; \|z^*\| \leq 1\} \subseteq \{C^*y^*; y^* \in D(C^*), \|y^*\| \leq k_2\}.$$

Moreover, in the equivalence (b) \Leftrightarrow (c) we can take $k_1 = k_2$.

LEMMA 3.2 (see [7]). *Let X, Y, Z be Banach spaces. Let $C : D(C) \subseteq X \rightarrow Y$ be linear, closed with dense domain, and let $F \in L(Z, Y)$. Let us state the conditions:*

- (i) $F(Z) \subset C(X)$;
- (ii) *there exists $k_1 > 0$ such that*

$$\{Fz; \|z\| \leq 1\} \subset \{Cx; x \in D(C), \|x\| \leq k_1\};$$

- (iii) *there exists $k_2 > 0$ such that*

$$\|F^*y^*\| \leq k_2 \|C^*y^*\|, \quad y^* \in D(C^*);$$

- (iv) *there exists $k_3 > 0$ such that*

$$\{Fz; \|z\| \leq 1\} \subseteq \text{cl} \{Cx; x \in D(C), \|x\| \leq k_3\}.$$

Then (i) \Leftrightarrow (ii), (iii) \Leftrightarrow (iv), with $k_2 = k_3$. If F is the identity operator, then (ii) \Leftrightarrow (iii) with $k_1 = k_2$.

Let us relate these abstract results to our problem (see also [7], [8], [10]). The control system (1.1) is null controllable at time t by L^p -controls, if $S(t)X \subseteq H(t)L^p(0, t; U)$. By virtue of Lemma 3.2, this is equivalent to the existence of $\alpha(t) > 0$ such that

$$(3.1) \quad \{S(t)x; \|x\| \leq \alpha(t)\} \subseteq \{H(t)u; u \in U_{\text{ad}}\},$$

which, in fact, means that $T(x) \leq t$ for $\|x\| \leq \alpha(t)$. We have thus proved the following result.

THEOREM 3.1. *Assume that U_{ad} is given by (2.1), $1 \leq p \leq \infty$. Then the control system (1.1) is null controllable at time t for every $t > 0$, by L^p -controls, if and only if the minimal time function is continuous in the origin of X .*

We also have the following corollary.

COROLLARY 3.1. *In the conditions of Theorem 3.1, assume further that the control system (1.1) is null controllable at every time $t > 0$, by L^p -controls. Then, V is open and the minimal time function is locally uniformly continuous on V . If, in addition, $1 < p \leq \infty$ and $S(t)$ is a contraction semigroup then $V = X$ and the minimal time function is uniformly continuous on X .*

Proof. The first part follows from Theorems 2.1 and 3.1. If $1 < p \leq \infty$ and $S(t)$ is a semigroup of contractions, then it follows by a result of Narukawa [19] that $V = X$. Finally, combining Theorem 3.1 and Corollary 2.1, we see that the minimal time function is uniformly continuous on X .

Remark 3.1. A more precise estimate for the function $\alpha(t)$ in (3.1) provides a corresponding estimate for the minimal time function around the origin (which here is considered as target). Along with Theorem 2.1, this provides a precise estimate for $T(\cdot)$ on the reachable set V . For example, suppose the function $\alpha : (0, \infty) \rightarrow (0, \infty)$ (or, at least, its restriction to $(0, \delta)$ for some $\delta > 0$) has an inverse. Then for x sufficiently near zero we take $t = \alpha^{-1}(\|x\|)$ and use (3.1). This implies $T(x) \leq \alpha^{-1}(\|x\|)$, for $\|x\|$ small. The problem lies in obtaining $\alpha(t)$ explicitly. One way might be by duality, using the equivalence (ii) \Leftrightarrow (iii) in Lemma 3.2, or (b) \Leftrightarrow (c) in Lemma 3.1. Note that Lemma 3.1 is useful in the case where the control process takes place in dual spaces, but this does not cover the situation when the control space is $L^1(0, t; U)$. More precisely, suppose that X and U are reflexive Banach spaces, $H(t) = C^*(t)$ and $1 < p \leq \infty$. Denoting $F(t) = S^*(t)$, if we provide an inequality of type

$$(3.2) \quad \omega(t) \|S^*(t)x^*\| \leq \|C(t)x^*\|, \quad x^* \in X^*,$$

then, using Lemma 3.1 we obtain (3.1) with $\alpha(t) = \rho\omega(t)$. See Example 4.2 for the illustration of this technique.

Another way to obtain the function $\alpha(t)$ in (3.1) is suggested by the work of Seidman [23]. Namely, let $C_t : X \rightarrow L^p(0, t; U)$ be the operator (initial data) \rightarrow (optimal null control for time t). Suppose that we have an estimate of type $\|C_t\| \leq \varphi(t)$ for t small. Then, we can take $\alpha(t) = \rho/\varphi(t)$ in (3.1). See Examples 4.3 and 4.4 for applications.

4. Examples. This section gives several applications of the abstract results and methods presented in §§ 2 and 3.

Example 4.1. Let us consider a control system described by the equation

$$(4.1) \quad y' = Ay + u,$$

where A generates a C_0 -semigroup, $S(t)$, $t \geq 0$, on a Banach space X . The mild solution

$$y(t, x, u) = S(t)x + \int_0^t S(t-s)u(s) ds, \quad t \geq 0,$$

is of type (1.1). We take

$$(4.2) \quad U_{\text{ad}} = \{u \in L^\infty(\mathbb{R}^+; X); \|u(t)\| \leq \rho \text{ a.e.}\}$$

and consider the target $x_1 \in X$.

THEOREM 4.1. *Let $x_1 \in D(A)$ be such that $\|Ax_1\| < \rho$. Let $M \geq 1$ and $\omega \geq 0$ be such that $\|S(t)\| \leq M \exp(\omega t)$. Then*

(a) *If $\omega > 0$, V is open and the minimal time function associated with (4.1), (4.2) is locally Lipschitz continuous on V .*

(b) *If $\omega = 0$, we have $V = X$, and the minimal time function associated with (4.1), (4.2) is Lipschitz continuous on X .*

Proof. Consider first $\omega > 0$. An easy calculation shows that if

$$\|x - x_1\| \leq \frac{\rho - \|Ax_1\|}{2\omega M},$$

then

$$\int_0^t \frac{1}{\|S(s)(x - x_1)\|} ds \geq \frac{1}{\rho - \|Ax_1\|}$$

for t sufficiently large.

Let $t > 0$ be such that we have equality and take the control

$$u(s) = -Ax_1 - \frac{\rho - \|Ax_1\|}{\|S(s)(x - x_1)\|} S(s)(x - x_1), \quad s \in [0, t].$$

Clearly, $u \in U_{ad}$ and $y(t, x, u) = x_1$. Hence, $x \in V(t)$ and therefore $T(x) \leq t$. We also have

$$(4.3) \quad \int_0^{T(x)} \frac{1}{\|S(s)(x - x_1)\|} dx \leq \frac{1}{\rho - \|Ax_1\|}.$$

This implies

$$T(x) \leq -\frac{1}{\omega} \log \frac{\rho - \|Ax_1\| - \omega M \|x - x_1\|}{\rho - \|Ax_1\|},$$

for all x that satisfies

$$\|x - x_1\| \leq \frac{\rho - \|Ax_1\|}{2\omega M}.$$

It is easy to prove now that the minimal time function is locally Lipschitz continuous in x_1 . By Theorem 2.1, it is locally Lipschitz continuous on V .

Note that if $S(s_0)(x - x_1) = 0$ for some s_0 , we can construct a continuous function $\varphi(\cdot)$ such that

$$\varphi(s) \leq 1/\|S(s)(x - x_1)\| \quad \text{for } s \geq 0$$

and such that $\int_0^\infty \varphi(s) ds = \infty$, and we work with $\varphi(\cdot)$ instead of $1/\|S(\cdot)(x - x_1)\|$.

In the case where $\omega = 0$, clearly, we have

$$\int_0^\infty \frac{ds}{\|S(s)(x - x_1)\|} = \infty \quad \text{for every } x \in X,$$

so that $V = X$ and

$$T(x) \leq \frac{M\|x - x_1\|}{\rho - \|Ax_1\|} \quad \text{for every } x \in X.$$

This shows that $T(\cdot)$ is Lipschitz continuous in x_1 , and hence, by Corollary 2.1, it is Lipschitz continuous on X .

Note that Corollary 2.1 was proved in the case where $\omega = 0$ and $M = 1$, but the proof also works in the case where $\omega = 0$ and every $M > 0$. This concludes the proof.

Remark 4.1. We have proved in [5], by another approach, a related result in the case where $x = 0$ and A (possibly) nonlinear.

Example 4.2. Consider the abstract wave equation

$$(4.4) \quad z'' = Az + u,$$

where A is a selfadjoint, positive definite linear operator defined on a dense domain $D(A)$ in a Hilbert space H . In addition, assume that A has a complete sequence of orthonormal eigenelements and a corresponding sequence $(\lambda_j)_{j \in \mathbb{N}}$ of real eigenvalues of finite multiplicity with $0 < \lambda_1 \leq \lambda_2 \leq \dots$ and $\lim_{j \rightarrow \infty} \lambda_j = \infty$. Let $V = D(A^{1/2})$, the domain of the square root of A , let $C(t) : H \rightarrow H(V \rightarrow V)$ and $S_1(t) : H \rightarrow V$ be the cosine and sine operators generated by A . See, e.g., [11], [14], and [15] for details.

For the initial conditions $z(0) = z_0$, $z'(0) = z_1$, $x = (z_0, z_1) \in V \times H$ and $u \in L^2_{\text{loc}}(R^+; H)$, the solution to system (4.4) is understood in the mild sense,

$$z(t, x, u) = C(t)z_0 + S_1(t)z_1 + \int_0^t S_1(t-s)u(s) ds,$$

and the derivation $z'(t, x, u)$ is given by

$$z'(t, x, u) = -AS_1(t)z_0 + C(t)z_1 + \int_0^t C(t-s)u(s) ds.$$

Let $y(t, x, u) = (z(t, x, u), z'(t, x, u)) \in V \times H$, $t \geq 0$. It is easy to see that $y(t, x, u)$ can be represented as in (1.1) in the space $V \times H$, where

$$S(t) = \begin{bmatrix} C(t) & S_1(t) \\ -AS_1(t) & C(t) \end{bmatrix}$$

and

$$H(t)u = \int_0^t S(t-s)Bu(s) ds,$$

with $B : H \rightarrow V \times H$ given by $Bu = \begin{pmatrix} 0 \\ u \end{pmatrix}$. (See Example 1.1.)

We study the minimal time function associated with (4.4) in two cases.

Case 1. The admissible set of controls is

$$(4.5) \quad U_{ad} = \{u \in L^\infty([0, \infty); H); \|u(t)\|_H \leq \rho \text{ a.e.}\}.$$

It is known that the control system (4.4) is null controllable at every time $t > 0$ by L^∞ -controls; that is, given some $t > 0$ and some initial state $x = (z_0, z_1) \in V \times H$, there exists a control function $u \in L^\infty([0, \infty); H)$ such that the corresponding mild solution $z(t, x, u)$ of (4.4) satisfies $z(t, x, u) = 0$ and $z'(t, x, u) = 0$ (see [4]). On the other hand, $S(t)$ is a contraction semigroup (in fact a group). Hence, by Corollary 3.1, the minimal time function associated with (4.4) and (4.5) is finite on $V \times H$ and uniformly continuous. We now give a more precise estimate.

THEOREM 4.2. *The minimal time function associated with (4.4) and (4.5) satisfies the estimate*

$$(4.6) \quad |T(x) - T(y)| \leq M_1 \|x - y\|^{1/3},$$

for every $x, y \in V \times H$ with $\|x - y\| \leq M_2$, where $M_1 = (30\sqrt{2}/\rho\lambda_1)^{1/3}$ and $M_2 = \sqrt{15/64}(\rho/\sqrt{\lambda_1})$. Here, for $x = (x_1, x_2) \in V \times H$, $\|x\| = (\|x_1\|^2 + \|x_2\|^2)^{1/2}$.

Proof. From the proof of Theorems 2.2 and 2.3 in [14] it follows that for every $x = (z_0, z_1) \in V \times H$ and $0 < t \leq \sqrt{15/8\lambda_1}$, there exists a control $u_t \in L^\infty(0, t; H)$ that transfers x to rest during $[0, t]$ and satisfies the estimate

$$\|u_t\| \leq \frac{30\sqrt{2}}{\lambda_1} t^{-3} \|x\|.$$

See [9] for more details. Note that a careful reading of [14] shows that this result is a consequence of an inequality of type (3.2). If $\|x\| \leq (t^3 \lambda_1 \rho)/(30\sqrt{2})$, then $u_t \in U_{\text{ad}}$, and thus, we obtain (3.1) with $\alpha(t) = (t^3 \lambda_1 \rho)/(30\sqrt{2})$ and $t \leq \sqrt{15/8\lambda_1}$. Taking $t = (30\sqrt{2}/\lambda_1 \rho)^{1/3} \|x\|^{1/3}$, we obtain

$$(4.7) \quad T(x) \leq M_1 \|x\|^{1/3} \quad \text{for } \|x\| \leq M_2.$$

See also Remark 3.1. We have just obtained an estimate for $T(\cdot)$ around zero. Since $S(t)$ is a group of contractions, (4.6) follows from (2.3) and (4.7).

Case 2. The admissible set of controls is

$$(4.8) \quad U_{\text{ad}} = \bigcup_{t>0} \{u \in L^2(0, t; H); \|u\|_2 \leq \rho\}.$$

THEOREM 4.3. *The minimal time function associated with (4.4) and (4.8) satisfies the estimate*

$$|T(x) - T(y)| \leq M_3 \|x - y\|^{2/3},$$

for every $x, y \in V \times H$ with $\|x - y\| \leq M_4$, where $M_3 = (30/\lambda_1)^{1/3} \rho^{-2/3}$ and $M_4 = \rho(15/8\lambda_1)^{1/3} (\lambda_1/30)^{3/4}$.

Proof. We apply again a result of [14] to deduce that for every $x = (z_1, z_2) \in V \times H$ and $0 < t \leq \sqrt{15/8\lambda_1}$ there exists a control $u_t \in L^2(0, t; H)$ that transfers x to zero within time t and satisfies the estimate

$$\|u_t\|_2 \leq 2 \left(\frac{15}{2\lambda_1} \right)^{1/2} t^{-3/2} \|x\|.$$

See also [9]. The proof is concluded in the same way as the proof of Theorem 4.2.

Example 4.3. Let Ω be a bounded domain in R^n whose boundary Γ is a C^∞ manifold. Let Δ denote the Laplacian operation on R^n and let the constants $a, b, a^2 + b^2 \neq 0, ab \geq 0$. We consider a control system described by the equation

$$(4.9) \quad \begin{aligned} y' - \Delta y &= 0 && \text{in } \Omega \times (0, \infty), \\ a \frac{\partial y}{\partial \nu} + by &= u && \text{in } \Gamma \times (0, \infty), \\ y(0) &= y_0 && \text{in } \Omega. \end{aligned}$$

Consider, first, the admissible set of controls

$$(4.10) \quad U_{\text{ad}} = \{u \in L^\infty(\Gamma \times (0, \infty)); |u(x, t)| \leq \rho \text{ a.e.}\}.$$

It is well known (see, e.g., [20]) that for $y_0 \in L^2(\Omega)$ and $u \in L^\infty(\Gamma \times (0, \infty))$, (4.9) has a unique weak solution $y(t, y_0, u)$ in $L^2(\Omega)$, which can be represented as in (1.1) such that (1.2) is verified. Clearly, in this case $p = \infty$, $U = L^\infty(\Gamma)$, and $X = L^2(\Omega)$. The semigroup $S(t)$ involved here is a C_0 -semigroup of contractions. On the other hand, in the case where $b \neq 0$, for every $t > 0$ the control system (4.9) is null controllable by $L^\infty(\Gamma \times (0, \infty))$ -controls (see [20]). Taking into account Corollary 3.1, we make the following deduction.

COROLLARY 4.1. *The minimal time function associated with (4.9) and (4.10), in the case where $b \neq 0$, is finite and uniformly continuous on $L^2(\Omega)$.*

We consider now the admissible set of controls

$$(4.11) \quad U_{\text{ad}} = \bigcup_{t>0} \{u \in L^2(0, t; L^2(\Gamma)), \|u\|_2 \leq \rho\}.$$

If $a \neq 0$, again the solution $y(t, y_0, u)$ is in $L^2(\Omega)$ for every $t > 0$ (see [21]), can be represented as in (1.1) with $H(t)$ continuous from $L^2(0, t; L^2(\Gamma))$ into $L^2(\Omega)$. Furthermore, we have the null controllability property for every $t > 0$ by L^2 -controls so we have (3.1). If $a = 0$, $H(t)u$ need not be in $L^2(\Omega)$ for every $u \in L^2(0, t; L^2(\Gamma))$. However, the operator $H(t) : D(H(t)) \rightarrow L^2(\Omega)$, where $D(H(t)) = \{u \in L^2(0, t; L^2(\Gamma)); H(t)u \in L^2(\Omega)\}$, is densely defined and closed (see [22, Lemma 2.15]); hence, we can apply Lemma 3.2 to deduce (3.1) from the null controllability property.

For $t > 0$, define the operator

$$C_t : L^2(\Omega) \rightarrow L^2(0, t; L^2(\Gamma)),$$

$$C_t y_0 = u_0,$$

where u_0 is the minimum norm control that transfers the initial state y_0 to zero within time t by controls in $L^2(0, t; L^2(\Gamma))$.

LEMMA 4.1 [23]. *There exists $\alpha > 0$, such that*

$$(4.12) \quad \|C_t\| \leq \exp(\alpha/t) \quad \text{for small } t.$$

We are now in position to prove the following theorem.

THEOREM 4.4. *The minimal time function associated with (4.9) and (4.11) satisfies the estimate*

$$(4.13) \quad |T(y_0) - T(z_0)| \leq \frac{k}{\log(\rho/\|y_0 - z_0\|)}$$

for some $k > 0$ and for every $y_0, z_0 \in L^2(\Omega)$ such that $\|y_0 - z_0\|$ is sufficiently small.

Proof. We first prove that

$$T(y_0) \leq \frac{\alpha}{\log(\rho/\|y_0\|)} \quad \text{for } \|y_0\| \text{ small,}$$

where α is given by Lemma 4.1.

Indeed, by (4.12), for small $t > 0$, every $y_0 \in L^2(\Omega)$ with $\|y_0\| < \rho$ can be transferred to zero by $u_0 \in L^2(0, t; L^2(\Gamma))$ with $\|u_0\| \leq \exp(\alpha/t)\|y_0\|$. Hence, if $\exp(\alpha/t)\|y_0\| \leq \rho$ then $y_0 \in V(t)$. Therefore, we have (3.1) with $\alpha(t) = \rho \exp(-\alpha/t)$. See also the final part of Remark 3.1. The inequality (4.14) now follows easily. To conclude the proof of our theorem we apply Theorem 2.1, more precisely, an inequality of type (2.4). Indeed,

$$\begin{aligned} |T(y_0) - T(z_0)| &\leq T(h(y_0, z_0)) \leq \frac{\alpha}{\log(\rho/\|h(y_0, z_0)\|)} \\ &\leq \frac{\alpha}{\log(\rho/\|y_0 - z_0\|)}. \end{aligned}$$

Here, $h(y_0, z_0) = \text{sign}(T(y_0) - T(z_0))S(\min\{T(y_0), T(z_0)\})(y_0 - z_0)$, and $\|h(y_0, z_0)\| \leq \|y_0 - z_0\|$ because $S(t)$ is a semigroup of contractions.

Remark 4.2. The estimate given in Theorem 4.4 seems to be optimal due to a result of Güichal [12], which states that the estimate (4.12) is optimal.

Example 4.4. Let us consider the boundary control problem of a vibrating plate

$$(4.15) \quad \begin{aligned} y'' + \Delta^2 y &= 0 && \text{on } \Omega \times (0, \infty), \\ \partial y / \partial \nu &= 0, \partial(\Delta y) / \partial \nu = u && \text{on } \Gamma \times (0, \infty), \\ y(0) &= y_0^1; y'(0) = y_0^2 && \text{on } \Omega, \end{aligned}$$

where $\Omega = (0, 1) \times (0, 1)$.

Resulting from [16], the control system (4.15) is null controllable for arbitrarily small time by L^2 -controls, with an estimate of type (4.12). Therefore, the minimal time function associated with the system (4.15) and the admissible set of controls (4.11) satisfies an estimate of type (4.13).

Example 4.5. Consider, finally, the wave equation with boundary control (Example 1.3). It is well known that the control system (1.4) is not null controllable for arbitrarily small time (see [18]). Therefore, by Theorem 3.1, the corresponding minimal time function is not continuous in zero.

Acknowledgment. I thank Professor Corneliu Ursescu for helpful discussions.

REFERENCES

- [1] V. BARBU, *The minimal time function for the nonlinear diffusion equation*, Libertas Mathematica, 10 (1990), pp. 123–130.
- [2] ———, *The dynamic programming equation for the time optimal problem in infinite dimension*, SIAM J. Control Optim., 29 (1991), pp. 445–456.
- [3] M. BARDI, *Boundary value problem for the minimum time function*, SIAM J. Control Optim., 27 (1989), pp. 776–785.
- [4] O. CĂRJĂ, *Local controllability of nonlinear evolution equations in Banach spaces*, Analele Stiintifice ale Univ. “Al.I. Cuza” Iași, Sect. I.a, Mat, 25 (1979), pp. 117–125.
- [5] ———, *On the minimal time function for distributed control systems in Banach spaces*, J. Optim. Theory Appl., 44 (1984), pp. 397–406.
- [6] ———, *On continuity of the minimal time function for distributed control systems*, Boll. Un. Mat. Ital., (6), 4-A (1985), pp. 293–302.
- [7] ———, *On constraint controllability of linear systems in Banach spaces*, J. Optim. Theory Appl., 56 (1988), pp. 215–225.
- [8] ———, *Range inclusion for convex processes on Banach spaces: applications in controllability*, Proc. Amer. Math. Soc., 105 (1989), pp. 185–191.
- [9] ———, *The minimal time function for vibrating systems*, in Differential Equations and Control Theory, V. Barbu, ed., Longman Scientific and Technical, 1991, pp. 58–62.
- [10] S. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.
- [11] H. O. FATTORINI, *The time optimal problem for distributed control of systems described by the wave equation*, in Control Theory of Systems Governed by Partial Differential Equations, A. K. Aziz, J. W. Wingate, M. J. Balas, eds., Academic Press, New York, San Francisco, London, 1977, pp. 305–320.
- [12] E. N. GUICHAL, *A lower bound of the norm of the control operator for the heat equation*, J. Math. Anal. Appl., 110 (1985), pp. 519–527.
- [13] O. HAJEK, *Geometric theory of time-optimal control*, SIAM J. Control Optim., 9 (1971), pp. 339–350.
- [14] W. KRABS, *On time-minimal distributed control of vibrating systems governed by an abstract wave equation*, Appl. Math. Optim., 13 (1985), pp. 137–149.
- [15] ———, *On time-minimal distributed control of vibrations*, Appl. Math. Optim., 19 (1989), pp. 65–73.
- [16] W. KRABS, G. LEUGERING AND T. SEIDMAN, *On boundary controllability of a vibrating plate*, Appl. Math. Optim., 13 (1985), pp. 205–229.
- [17] E. B. LEE AND L. MARCUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [18] J. L. LIONS, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués, Tome I: Contrôlabilité exacte*, Masson, Paris, 1988.

- [19] K. NARUKAWA, *Admissible null controllability and optimal time control*, Hiroshima Math. J., 11 (1981), pp. 533–551.
- [20] G. SCHMIDT, *The “bang-bang” principle for the time-optimal problem in boundary control of the heat equation*, SIAM J. Control Optim., 18 (1980), pp. 101–107.
- [21] T. SEIDMAN, *Observation and prediction for the heat equation. IV: Patch observability and controllability*, SIAM J. Control Optim., 15 (1977), pp. 412–427.
- [22] ———, *Regularity of optimal boundary control for parabolic equations. I: Analyticity*, SIAM J. Control Optim., 20 (1982), pp. 428–453.
- [23] ———, *Two results on exact boundary control of parabolic equations*, Appl. Math. Optim., 11 (1984), pp. 145–152.
- [24] R. TRIGGIANI, *Exact boundary controllability on $L_2(\Omega) \times H^{-1}(\Omega)$ of the wave equation with Dirichlet boundary control acting on a portion of the boundary $\partial\Omega$, and related problems*, Appl. Math. Optim., 18 (1988), pp. 241–277.
- [25] G. WEISS, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

A STATE-SPACE ALGORITHM FOR THE SOLUTION OF THE 2-BLOCK SUPEROPTIMAL DISTANCE PROBLEM*

I. M. JAIMOUKHA† AND D. J. N. LIMEBEER†

Abstract. A state-space algorithm for computing the solution of the 2-block superoptimal distance problem (SODP) is presented. Given a rational and antistable matrix function $R(s) = [R_{11}(s) \ R_{12}(s)]$, find all stable approximations $Q(s)$ that lexicographically minimize the singular values of the error function $E(s) = [R_{11}(s) \ R_{12}(s) + Q(s)]$. Conditions are given for which the superoptimal approximation is unique. In addition, an a priori upper bound on the MacMillan degree of the approximation is given. The algorithm may be stopped after minimizing a given number of singular values. This premature termination of the algorithm carries with it an expected saving in the computational effort and a predictable reduction in the MacMillan degree of the approximation. The algorithm only requires standard linear algebraic computations and is, therefore, easily implemented.

Key words. superoptimal general distance problem, 2-block general distance problem, H_∞ -optimal control

AMS subject classification. 93C35

1. Notation and definitions.

$\mathbb{R}, \mathbb{R}_+, \mathbb{C}$	real, nonnegative and complex numbers
$\mathbb{C}_+(\bar{\mathbb{C}}_+), \mathbb{C}_-(\bar{\mathbb{C}}_-)$	open (respectively, closed) right half plane, open (respectively, closed) left half-plane
$\lambda(A), \lambda_{\max}(A)$	eigenvalue of square matrix A , largest eigenvalue of A
A^*	complex conjugate transpose of $A \in \mathbb{C}^{m \times l}$
$A \geq 0, A > 0$	A is positive semidefinite (respectively, positive definite)
$\mathcal{L}_\infty^{p \times m}$	space of $p \times m$ matrix functions with entries bounded on the $j\omega$ axis
$\ (\cdot)\ _\infty$	\mathcal{L}_∞ -norm of matrices in \mathcal{L}_∞
$\mathcal{H}_\infty^{+p \times m}, \mathcal{H}_\infty^{-p \times m}$	subspaces of $\mathcal{L}_\infty^{p \times m}$; matrices that are analytic in $\bar{\mathbb{C}}_+$ (respectively, $\bar{\mathbb{C}}_-$)
$[G(s)]^*, [G(s)]^{-*}$	$G(-\bar{s})^*$, the para-Hermitian conjugate of $G(s)$, $[G^*(s)]^{-1}$
$\deg [G(s)]$	MacMillan degree of the rational function $G(s)$
$[G(s)]_+, [G(s)]_-$	stable (respectively, antistable) projection of $G(s)$
Prefix \mathbb{R}	denotes real rational.

Associated with every rational transfer function matrix $G(s)$ is a *state-space realization* $G(s) = D + C(sI - A)^{-1}B$, where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times l}$, $C \in \mathbb{R}^{m \times n}$ and $D \in \mathbb{R}^{m \times l}$, and where $n \geq \deg [G(s)]$. The alternative notation $G(s) = (A, B, C, D)$ or

$$G(s) \stackrel{s}{=} \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

is also used. Occasionally we write

$$\begin{bmatrix} R_{11}(s) & R_{12}(s) \\ R_{21}(s) & R_{22}(s) \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array} \right],$$

* Received by the editors July 15, 1991; accepted for publication (in revised form) March 12, 1992. This research was supported in part by the Arab British Chamber Charitable Foundation Scholarship, and in part by the Overseas Research Studentship Award Scheme.

† Centre for Process Systems Engineering and Department of Electrical Engineering, Imperial College, Exhibition Road, London SW7 2BY.

in which $R_{ij}(s) \stackrel{s}{=} (A, B_j, C_i, D_{ij})$ for $i, j = 1, 2$. If D is nonsingular, then

$$(1) \quad [G(s)]^{-1} \stackrel{s}{=} (A - BD^{-1}C, -BD^{-1}, D^{-1}C, D^{-1}).$$

If $[G(s)]^{-1} = [G(s)]^*$, then $G(s)$ is said to be an *all-pass* system and satisfies $[G(s)] \cdot [G(s)]^* = [G(s)]^* [G(s)] = I$. Occasionally, we say that $G(s)$ is all-pass for some $\gamma \in \mathbb{R}_+$. This is taken to mean that $\gamma^{-1}G(s)$ is all-pass. The rational function $G(s)$ is called *stable* if it has no poles in $\bar{\mathbb{C}}_+$. If $G(s)$ has no poles in $\bar{\mathbb{C}}_-$ it is called *antistable*. If $G(s) \in \mathcal{H}_\infty^+$ has $\|G\|_\infty \leq 1$, it is called a *stable contraction*. The set of all stable contractions is denoted \mathcal{BH}_∞^+ . We say that $G(s) \in \gamma^{-1}\mathcal{BH}_\infty^+$ if $\gamma G(s) \in \mathcal{BH}_\infty^+$. If a *basis change* T is introduced into a state-space realization of $G(s)$, this is taken to mean the similarity transformation

$$(2) \quad G(s) \stackrel{s}{=} (A, B, C, D) \xrightarrow{T} G(s) \stackrel{s}{=} (T^{-1}AT, T^{-1}B, CT, D).$$

In most cases, $G(s)$ will be abbreviated to G .

If $U \in \mathcal{L}_\infty^{l \times q}$ and

$$H = \begin{matrix} & m & l \\ p & \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} & \end{matrix} \in \mathcal{L}_\infty^{(p+q) \times (m+l)}$$

we define the *lower linear fractional map* (LLFM) $\mathcal{F}_l(H, U) := H_{11} + H_{12}U(I - H_{22}U)^{-1}H_{21}$ provided $[I - H_{22}(\infty)U(\infty)]$ is invertible. If $U \in \mathcal{L}_\infty^{m \times p}$ we define the *upper linear fractional map* (ULFM) $\mathcal{F}_u(H, U) := H_{22} + H_{21}U(I - H_{11}U)^{-1}H_{12}$ provided $[I - H_{11}(\infty)U(\infty)]$ is invertible. If \mathcal{U} is a set, then $\mathcal{F}_l(H, \mathcal{U})$ is taken to denote the set $\{\mathcal{F}_l(H, U) : U \in \mathcal{U}\}$. If $G_1, G_2, G_3 \in \mathcal{L}_\infty$ have appropriate dimensions, and \mathcal{U} is a set, then $G_1 + G_2\mathcal{U}G_3$ is taken to denote the set $\{G_1 + G_2UG_3 : U \in \mathcal{U}\}$.

A problem of presenting the superoptimal algorithm is the elaborate notation. To simplify the presentation, we adopt the following special notation and definitions. If $A \in \mathbb{C}^{p \times m}$, then $\sigma_i(A)$ denotes the i th largest singular value of A . If $G \in \mathcal{L}_\infty^{p \times m}$, then

$$s_i^\infty(G) := \sup_{\omega \in \mathbb{R}} \sigma_i[G(j\omega)]$$

denotes the supremum of the i th largest singular value of G over the imaginary axis (including ∞). Clearly, $s_1^\infty(G) = \|G\|_\infty$. If $R \in \mathbb{R}\mathcal{H}_\infty^{-p \times (l+m)}$ is partitioned as

$$(3) \quad R = \begin{matrix} & l & m \\ p & \begin{bmatrix} R_{11} & R_{12} \end{bmatrix} & \end{matrix} \in \mathbb{R}\mathcal{H}_\infty^{-p \times (l+m)},$$

we define the optimal level of R as

$$s_1(R) := \inf_{Q \in \mathcal{H}_\infty^{+p \times m}} s_1^\infty([R_{11} \quad R_{12} + Q]),$$

and the set of all optimal approximations of R as

$$\mathcal{S}_1(R) := \{Q \in \mathcal{H}_\infty^{+p \times m} : s_1^\infty([R_{11} \quad R_{12} + Q]) = s_1\}.$$

If p and m are both greater than 1, then we define the first and subsequent superoptimal levels of R as

$$s_i(R) := \inf_{Q \in \mathcal{S}_{i-1}(R)} s_i^\infty([R_{11} \quad R_{12} + Q]) \quad i = 1, 2, \dots,$$

and the set of all i th level superoptimal approximations of R as

$$\mathcal{S}_i(R) := \{Q \in \mathcal{S}_{i-1}(R) : s_i^\infty([R_{11} \quad R_{12} + Q]) = s_i(R)\} \quad i = 1, 2, \dots,$$

in which $s_0 = \infty$ and $S_0(R) = \mathcal{H}_{\infty}^{+p \times m}$. Clearly, the optimal level is equal to the first superoptimal level, and the set of all optimal approximations is equal to the set of all first-level superoptimal approximations. If $\gamma > s_1(R)$, then γ is said to be a suboptimal level of R . The set of all suboptimal approximations of R at level γ is defined as

$$\mathcal{S}(R, \gamma) := \{Q \in \mathcal{H}_{\infty}^{+p \times m} : s_1^{\infty}([R_{11} \quad R_{12} + Q]) \leq \gamma\}.$$

2. Introduction. It is well known [4], [14] that a large class of \mathcal{H}_{∞} -control problems may be reduced to the 2-block optimal distance problem via the Youla parametrization of all stabilizing controllers [18]. In the 2-block optimal distance problem, we are given $R \in \mathbb{R}\mathcal{H}_{\infty}^{-p \times (l+m)}$ partitioned as (3), and we seek to find the optimal level $s_1(R)$ and the set $S_1(R)$ of all optimal approximations of R . In control problems, R is a function of a plant model and various weighting functions [4], [14]. In general, the solution of the 2-block optimal distance problem is hardly ever unique [4], [14]. The question then arises as to whether any of these optimal solutions is best in some sense. One way of recovering uniqueness is to strengthen the optimality requirement. Specifically, we request that the second and subsequent singular values are minimized with respect to lexicographic ordering [19].

Problem 1. Suppose we are given a rational antistable matrix R partitioned as in (3). Then for $i = 1, 2, \dots$ we are required to find the superoptimal levels $s_1(R), \dots, s_i(R)$ and the set $S_i(R)$ of all i th level superoptimal approximations of R , where we set $s_0 = \infty$ and $S_0(R) = \mathcal{H}_{\infty}^{+p \times m}$.

When $R_{11} = 0$, Problem 1 reduces to the 1-block SODP first proposed by Young [19], who showed, using operator theoretic techniques, that the superoptimal approximation is unique [19]. State-space algorithms for calculating the solution are given in [9], [15], and [17], where it is shown that the superoptimal approximation has a surprisingly low MacMillan degree. Gu, Tsai, and Postlethwaite [8] give a solution to the 2-block SODP using a technique that is different from the one used in this paper. However, their algorithm is unnecessarily complicated and cannot be readily generalized to the solution of the 4-block SODP. In contrast, we give a new and simple algorithm for the solution of both the 1-block and 2-block SODPs, using essentially the same technique. More precisely, we prove that, under certain conditions, it is possible to construct a sequence of 2-block systems R_1, \dots, R_k of the same form as R such that $s_i(R) = s_1(R_i)$ and $S_i(R) = S_{i-1}[S_1(R_i)]$, where S_1, \dots, S_{k-1} is a sequence of operators to be defined later. Thus, the solution of the i th level SODP is reduced to the solution of i optimal problems. We also prove that there exist i th level superoptimal approximations of R of MacMillan degree $\leq (n-1) + (n-2) + \dots + (n-i)$, where $n = \deg(R)$. Furthermore, sufficient conditions are given for which the superoptimal approximation of R is unique. Finally, we demonstrate that the algorithm can be readily extended to the solution of the 4-block SODP.

The paper is laid out as follows. Section 3 gives a review of the 2-block optimal distance problem. Section 4 gives the solution of the second-level superoptimal problem and § 5 gives a cancellation analysis of this solution. Section 6 extends the solution to the full superoptimal problem and § 7 outlines the solution to the 4-block superoptimal problem. A few examples are given in § 8, and finally, the conclusions are given in § 9.

3. The 2-block optimal distance problem. This section outlines the solution of the 2-block optimal distance problem: for $R \in \mathbb{R}\mathcal{H}_{\infty}^{-p \times (l+m)}$ given by the state-space realization

$$(4) \quad R = \begin{bmatrix} R_{11} & R_{12} \end{bmatrix} \stackrel{l}{=} \begin{bmatrix} A & B_1 & B_2 \\ C_1 & 0 & 0 \end{bmatrix} \in \mathbb{R}\mathcal{H}_{\infty}^{-p \times (l+m)},$$

where

$$(5) \quad A \in \mathbb{R}^{n \times n}, \quad \operatorname{Re} [\lambda_i(A)] > 0 \quad \forall i,$$

find the optimal level $s_1(R)$ and the set $\mathbb{S}_1(R)$ of all optimal approximations of R . The solution to this problem has been developed in [2], [7], and [13]. We assume that the optimal level $s_1(R)$ can be calculated to any desired degree of accuracy using the γ -iteration [4], [7], and furthermore that

$$(6) \quad s_1(R) > \|R_{11}\|_\infty.$$

The following theorem gives the set $\mathbb{S}_1(R)$ in the form of a linear fractional map.

THEOREM 1 (see [2], [7], [13]). *Let $s_1 := s_1(R)$ and assume that (5) and (6) are satisfied. Then*

(i) *There exists an embedding of R of the form*

$$H := \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{array}{c} p \\ l \\ m-1 \end{array} \left[\begin{array}{cc|c} R_{11} & R_{12} + Q_{12} & Q_{13} \\ R_{21} & R_{22} + Q_{22} & Q_{23} \\ 0 & Q_{32} & Q_{33} \end{array} \right] \begin{array}{c} l \\ m \\ p-1 \end{array}$$

such that $HH^* = H^*H = s_1^2 I$; $R_{ij}, (R_{21})^{-1} \in \mathbb{RH}_\infty^-, i, j = 1, 2$; $Q_{ij} \in \mathbb{RH}_\infty^+, i = 1, 2, 3$; $j = 2, 3$ and

$$(7) \quad \|H_{22}\|_\infty < s_1.$$

Furthermore,

$$\mathbb{S}_1(R) = \mathcal{F}_1(Q_a, s_1^{-1} \mathcal{BH}_\infty^{(p-1) \times (m-1)})$$

and

$$[R_{11} \quad R_{12} + \mathbb{S}_1(R)] = \mathcal{F}_l(H, \mathcal{U}_1)$$

where

$$\mathcal{U}_1 := \{[0_{(p-1) \times l} \quad u] : u \in s_1^{-1} \mathcal{BH}_\infty^{(p-1) \times (m-1)}\}, \quad Q_a = \begin{bmatrix} Q_{12} & Q_{13} \\ Q_{32} & Q_{33} \end{bmatrix}.$$

(ii) *There exist real matrices*

$$\begin{aligned} A_H &= \begin{array}{c} n \\ n-1 \end{array} \begin{bmatrix} A & 0 \\ 0 & A_Q \end{bmatrix}, & B_H &= \begin{array}{c} l \\ n-1 \end{array} \begin{bmatrix} B_1 & B_2 & 0 \\ 0 & B_{Q_2} & B_{Q_3} \end{bmatrix}, \\ C_H &= \begin{array}{c} p \\ l \\ m-1 \end{array} \begin{bmatrix} C_1 & C_{Q_1} \\ C_2 & C_{Q_2} \\ 0 & C_{Q_3} \end{bmatrix}, & D_H &= \begin{array}{c} l \\ p \\ l \\ m-1 \end{array} \begin{bmatrix} 0 & D_{12} & D_{13} \\ s_1 I & 0 & 0 \\ 0 & D_{32} & 0 \end{bmatrix}, \\ P_H &= \begin{array}{c} n \\ n-1 \end{array} \begin{bmatrix} P_1 & P_3 \\ P_3^* & P_2 \end{bmatrix}, & Q_H &= \begin{array}{c} n \\ n-1 \end{array} \begin{bmatrix} Q_1 & Q_3 \\ Q_3^* & Q_2 \end{bmatrix} \end{aligned}$$

such that

$$(8) \quad A_Q \in \mathbb{R}^{(n-1) \times (n-1)}, \quad \operatorname{Re} [\lambda_i(A_Q)] < 0 \quad \forall i$$

$$P_H Q_H = Q_H P_H = s_1^2 I, \quad D_H D_H^* = D_H^* D_H = s_1^2 I, \quad A_H^* Q_H + Q_H A_H + C_H^* C_H = 0,$$

$$(9) \quad D_H^* C_H + B_H^* Q_H = 0, \quad D_H B_H^* + C_H P_H = 0, \quad A_H P_H + P_H A_H^* + B_H B_H^* = 0.$$

Furthermore, H and Q_a have state-space realizations given by

$$(10) \quad H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \stackrel{s}{=} \begin{bmatrix} A_H & B_H \\ C_H & D_H \end{bmatrix}, \quad Q_a := \begin{bmatrix} Q_{12} & Q_{13} \\ Q_{32} & Q_{33} \end{bmatrix} \stackrel{s}{=} \begin{bmatrix} A_Q & B_{Q_2} & B_{Q_3} \\ C_{Q_1} & D_{12} & D_{13} \\ C_{Q_3} & D_{32} & 0 \end{bmatrix}.$$

Remark 1. In the interest of a clear presentation the optimal level s_1 is taken to have multiplicity one. Relaxing this assumption will lead to a messy indexing system without introducing any serious technical difficulties. See [7] for a full treatment of the general case for the optimal distance problem. The assumption in (4) that $R(\infty) = 0$ is used to simplify the notation and can be removed with minor modifications to the algorithm [7].

Remark 2. If $\min(p, m) = 1$, then the optimal, and hence, the superoptimal approximation of R is unique and given by Q_{12} . If $n = 1$, then Q_{12} is a constant equal to D_{12} . Using the results in [7] it can be shown that the error $[R_{11} \quad R_{12} + D_{12}]$ has rank one so that D_{12} is the superoptimal solution.

Remark 3. In this paper, we will not give an explicit construction of the optimal generator H . However, we make extensive use of its properties as summarized in Theorem 1. For a full construction of the optimal generator H , see [7, Thm. 4.4].

4. The second-level SODP. This section considers the solution of the second-level SODP.

Problem 2. Suppose we are given a rational antistable matrix R partitioned as in (4). Then we are required to find the second superoptimal level $s_2(R)$ and the set $\mathbb{S}_2(R)$ of all second-level superoptimal approximations of R .

In the following theorem we derive the solution of the second level SODP associated with R using H only. Since $HH^* = H^*H = s_1^2 I$, we can regard H as an all-pass embedding of H_{22} . The next theorem uses standard spectral factorization results to give a construction of another embedding \bar{H} of H_{22} , which also satisfies $\bar{H}\bar{H}^* = \bar{H}^*\bar{H} = s_1^2 I$.

THEOREM 2. Let H be the generator of all optimal error functions of R given by Theorem 1. Then (i) There exists an embedding of H_{22} of the form

$$(11) \quad \bar{H} = \begin{bmatrix} \bar{H}_{11} & \bar{H}_{12} \\ \bar{H}_{21} & \bar{H}_{22} \end{bmatrix} = \begin{matrix} l & m-1 & p-1 \\ p-1 & & \\ & l & \\ m-1 & & \end{matrix} \begin{bmatrix} \bar{R}_{11} & \hat{R}_{12} + \bar{Q}_{12} & \bar{Q}_{13} \\ \bar{R}_{21} & \bar{R}_{22} + \bar{Q}_{22} & \bar{Q}_{23} \\ 0 & \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix}$$

such that

$$(12) \quad \bar{H}^* \bar{H} = \bar{H} \bar{H}^* = s_1^2 I,$$

$$(13) \quad \bar{R}_{21}^{-1}, \bar{R}_{21}, \bar{R}_{22}, \hat{R} := [\hat{R}_{11} \quad \hat{R}_{12}] \in \mathbb{R}\mathcal{H}_{\infty}^{-}, \quad \bar{Q}_{13}^{-1}, \bar{Q}_{32}^{-1}, \bar{Q}_{12}, \bar{Q}_{13}, \bar{Q}_{22}, \bar{Q}_{32} \in \mathbb{R}\mathcal{H}_{\infty}^{+}.$$

(ii) The set of all suboptimal approximations of \hat{R} is given by

$$(14) \quad \mathbb{S}(\hat{R}, s_1) = \mathcal{F}_l(\bar{Q}_a, s_1^{-1} \mathcal{B}\mathcal{H}_{\infty}^{+(p-1) \times (m-1)})$$

and the corresponding set of all suboptimal error functions of \hat{R} is given by

$$(15) \quad [\hat{R}_{11} \quad \hat{R}_{12} + \mathbb{S}(\hat{R}, s_1)] = \mathcal{F}_l(\bar{H}, \mathcal{U}_1),$$

where \mathcal{U}_1 is defined in Theorem 1 and

$$\bar{Q}_a := \begin{bmatrix} \bar{Q}_{12} & \bar{Q}_{13} \\ \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix}.$$

Proof. (i) Using (7), it is a simple exercise to give a construction of \bar{H} using standard spectral factorization theory [1] using either transfer function or state-space techniques. In this paper, we give a state-space construction of \bar{H} . It follows from Theorem 1 that H_{22} has the state-space realization

$$H_{22} \stackrel{s}{=} \left[\begin{array}{c|c} A_Q & B_{Q_3} \\ \hline C_{Q_2} & 0 \\ C_{Q_3} & 0 \end{array} \right].$$

Hence, it follows from (7) that there exist stabilizing solutions \bar{P}_2 and \bar{Q}_2 to the algebraic Riccati equations

$$(16) \quad \begin{aligned} A_Q \bar{P}_2 + \bar{P}_2 A_Q^* + B_{Q_3} B_{Q_3}^* + s_1^{-2} \bar{P}_2 C_{Q_3}^* C_{Q_3} \bar{P}_2 &= 0, \\ A_Q^* \bar{Q}_2 + \bar{Q}_2 A_Q + C_{Q_2}^* C_{Q_2} + C_{Q_3}^* C_{Q_3} + s_1^{-2} \bar{Q}_2 B_{Q_3} B_{Q_3}^* \bar{Q}_2 &= 0, \end{aligned}$$

respectively [2]. Since \bar{P}_2 and \bar{Q}_2 are stabilizing,

$$(17) \quad \operatorname{Re} [\lambda_i (A_Q + s_1^{-2} \bar{P}_2 C_{Q_3}^* C_{Q_3})] < 0 \quad \forall i, \quad \operatorname{Re} [\lambda_i (A_Q + s_1^{-2} B_{Q_3} B_{Q_3}^* \bar{Q}_2)] < 0 \quad \forall i.$$

Define

$$\bar{R} := \bar{Q}_2 \bar{P}_2 - s_1^2 I.$$

Then using the fact that \bar{P}_2 and \bar{Q}_2 are continuous and decreasing functions of s_1 [10], it is easy to show that (7) implies that $\lambda_{\max}(\bar{P}_2 \bar{Q}_2) < s_1^2$ [11], from which it follows that \bar{R} is invertible. Defining

$$(18) \quad \bar{H} = \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} + \bar{Q}_{12} & \bar{Q}_{13} \\ \bar{R}_{21} & \bar{R}_{22} + \bar{Q}_{22} & \bar{Q}_{23} \\ 0 & \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{cc|cc} \hat{A} & 0 & \hat{B}_1 & \hat{B}_2 & 0 \\ 0 & A_Q & 0 & \bar{B}_{Q_2} & B_{Q_3} \\ \hline \hat{C}_1 & \bar{C}_{Q_1} & 0 & 0 & s_1 I \\ \bar{C}_2 & C_{Q_2} & s_1 I & 0 & 0 \\ 0 & C_{Q_3} & 0 & s_1 I & 0 \end{array} \right] \stackrel{s}{=} \left[\begin{array}{c|c} A_{\bar{H}} & B_{\bar{H}} \\ \hline C_{\bar{H}} & D_{\bar{H}} \end{array} \right]$$

in which

$$(19) \quad \bar{Q}_a := \begin{bmatrix} \bar{Q}_{12} & \bar{Q}_{13} \\ \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|cc} A_Q & \bar{B}_{Q_2} & B_{Q_3} \\ \hline \bar{C}_{Q_1} & 0 & s_1 I \\ C_{Q_3} & s_1 I & 0 \end{array} \right],$$

and

$$(20) \quad \hat{R} = {}_{p-1} \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|c} \hat{A} & \begin{matrix} \hat{B}_1 & \hat{B}_2 \end{matrix} \\ \hline \hat{C}_1 & \begin{matrix} 0 & 0 \end{matrix} \end{array} \right] \in \mathbb{R} \mathcal{H}_{\infty}^{-(p-1) \times (l+m-1)},$$

$$(21) \quad \begin{aligned} \hat{A} &:= -(A_Q + s_1^{-2} \bar{P}_2 C_{Q_3}^* C_{Q_3})^*, \\ \bar{C}_{Q_1} &:= -s_1^{-1} B_{Q_3}^* \bar{Q}_2, \quad \hat{B}_1 := s_1 \bar{R}^{-1} C_{Q_2}^*, \quad \hat{B}_2 := -s_1^{-1} C_{Q_3}^*, \\ \hat{C}_1 &:= s_1^{-1} B_{Q_3}^* \bar{R}, \quad \bar{C}_2 := -s_1^{-1} \hat{B}_1^* \bar{Q}_1, \quad \bar{B}_{Q_2} := -s_1^{-1} \bar{P}_2 C_{Q_3}^*. \end{aligned}$$

Then

$$(22) \quad \hat{A} \in \mathbb{R}^{(n-1) \times (n-1)}$$

since $A_Q \in \mathbb{R}^{(n-1) \times (n-1)}$. Furthermore, (17) implies that

$$(23) \quad \operatorname{Re} [\lambda_i(\hat{A})] > 0 \quad \forall i.$$

Equations (16) and definitions (21) are used to verify that the realization of \bar{H} given in (18) satisfies the all-pass equations

$$\begin{aligned} A_{\bar{H}}^* Q_{\bar{H}} + Q_{\bar{H}} A_{\bar{H}} + C_{\bar{H}}^* C_{\bar{H}} &= 0, \quad D_{\bar{H}}^* C_{\bar{H}} + B_{\bar{H}}^* Q_{\bar{H}} = 0, \quad D_{\bar{H}} D_{\bar{H}}^* = D_{\bar{H}}^* D_{\bar{H}} = s_1^2 I, \\ A_{\bar{H}} P_{\bar{H}} + P_{\bar{H}} A_{\bar{H}}^* + B_{\bar{H}} B_{\bar{H}}^* &= 0, \quad D_{\bar{H}} B_{\bar{H}}^* + C_{\bar{H}} P_{\bar{H}} = 0, \quad P_{\bar{H}} Q_{\bar{H}} = Q_{\bar{H}} P_{\bar{H}} = s_1^2 I \end{aligned}$$

with

$$P_{\bar{H}} = \begin{bmatrix} \hat{P}_1 & I \\ I & \bar{P}_2 \end{bmatrix} := \begin{bmatrix} \bar{Q}_2 \bar{R}^{-*} & I \\ I & \bar{P}_2 \end{bmatrix}, \quad Q_{\bar{H}} = \begin{bmatrix} \bar{Q}_1 & -\bar{R}^* \\ -\bar{R} & \bar{Q}_2 \end{bmatrix} := \begin{bmatrix} \bar{P}_2 \bar{R} & -\bar{R}^* \\ -\bar{R} & \bar{Q}_2 \end{bmatrix}.$$

Hence, \bar{H} is all-pass at s_1 from [5, Thm. 5.1] and this proves (12). The state-space realizations (18)–(20), definitions (21), and (16) yield the following identities

$$\begin{aligned} \bar{Q}_{13}^{-1} &\stackrel{s}{=} (A_Q - s_1^{-1} B_{Q_3} C_{Q_1}, -s_1^{-1} B_{Q_3}, s_1^{-1} C_{Q_1}, s_1^{-1} I) \\ &\stackrel{s}{=} (A_Q + s_1^{-2} B_{Q_3} B_{Q_3}^* \bar{Q}_2, -s_1^{-1} B_{Q_3}, s_1^{-1} \bar{C}_{Q_1}, s_1^{-1} I), \\ \bar{Q}_{32}^{-1} &\stackrel{s}{=} (A_Q - s_1^{-1} B_{Q_2} C_{Q_3}, -s_1^{-1} B_{Q_2}, s_1^{-1} C_{Q_3}, s_1^{-1} I) \\ &\stackrel{s}{=} (A_Q + s_1^{-2} \bar{P}_2 C_{Q_3}^* C_{Q_3}, -s_1^{-1} \bar{B}_{Q_2}, s_1^{-1} C_{Q_3}, s_1^{-1} I), \\ \bar{R}_{21}^{-1} &\stackrel{s}{=} (A - s_1^{-1} \hat{B}_1 \bar{C}_2, -s_1^{-1} \hat{B}_1, s_1^{-1} \bar{C}_2, s_1^{-1} I) \\ &\stackrel{s}{=} [-\bar{R}^{-1} (A_Q + s_1^{-2} B_{Q_3} B_{Q_3}^* \bar{Q}_2)^* \bar{R}, -s_1^{-1} \hat{B}_1, s_1^{-1} \bar{C}_2, s_1^{-1} I]. \end{aligned}$$

Hence, (17) and (23) prove (13).

(ii) The fact that $s_1(R)$ is a suboptimal level of \hat{R} follows from part (i) since $\bar{Q}_{12} \in \mathbb{R} \mathcal{H}_{\infty}^+$ and

$$\|\bar{H}_{11}\|_{\infty} = \|\begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} + \bar{Q}_{12} \end{bmatrix}\|_{\infty} \leq \|\bar{H}\|_{\infty} = s_1(R).$$

Finally, (14) and (15) follow from (i) and [7, Thm. 3.6]. \square

The next theorem, which is the main result, shows that the solution of the second-level SODP associated with R is equivalent to the solution of the optimal problem associated with \hat{R} .

THEOREM 3. *Let H and \bar{H} be as given in Theorems 1 and 2, respectively. Then*

$$(24) \quad s_2(R) = s_1(\hat{R})$$

and

$$(25) \quad \mathbb{S}_2(R) = S[\mathbb{S}_1(\hat{R})],$$

where

$$S(\cdot) = \mathcal{F}_l\{Q_a, \mathcal{F}_u[\bar{Q}_a^{-1}, (\cdot)]\},$$

and where Q_a and \bar{Q}_a are defined in Theorems 1 and 2, respectively.

Proof. Since $HH^* = H^*H = s_1^2 I$ and $\bar{H}\bar{H}^* = \bar{H}^*\bar{H} = s_1^2 I$, it follows that

$$(26) \quad H_{11}H_{21}^* = -H_{21}H_{22}^*, \quad \bar{H}_{11} = -\bar{H}_{12}H_{22}^*\bar{H}_{21}^*,$$

$$(27) \quad \bar{H}_{21}\bar{H}_{21}^* = s_1^2 I - H_{22}H_{22}^* = H_{21}H_{21}^*, \quad \bar{H}_{12}^*\bar{H}_{12} = s_1^2 I - H_{22}^*H_{22} = H_{12}^*H_{12}.$$

Define

$$(28) \quad V_\perp := H_{12}\bar{H}_{12}^{-1}, \quad W_\perp := H_{21}^*\bar{H}_{21}^*.$$

Then (27) implies that

$$V_\perp^*V_\perp = I_{p-1}, \quad W_\perp^*W_\perp = I_{l+m-1}.$$

Hence, there exist

$$V = [v \quad V_\perp] \in \mathbb{R}\mathcal{L}_\infty^{p \times p}, \quad W = [w \quad W_\perp] \in \mathbb{R}\mathcal{L}_\infty^{(l+m) \times (l+m)}$$

such that V and W are all-pass [5]. Using (26) and definition (28) we obtain

$$V_\perp^*H_{12} = \bar{H}_{12}^*H_{12}^*H_{12} = \bar{H}_{12}^*\bar{H}_{12}^*\bar{H}_{12} = \bar{H}_{12},$$

$$H_{21}W_\perp = H_{21}H_{21}^*\bar{H}_{21}^* = \bar{H}_{21}\bar{H}_{21}^*\bar{H}_{21}^* = \bar{H}_{21},$$

$$V_\perp^*H_{11}W_\perp = V_\perp^*H_{11}H_{21}^*\bar{H}_{21}^* = -V_\perp^*H_{12}H_{22}^*\bar{H}_{21}^* = -\bar{H}_{12}H_{22}^*\bar{H}_{21}^* = \bar{H}_{11}.$$

It follows that

$$(29) \quad \begin{bmatrix} V^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} v^*H_{11}w & v^*H_{11}W_\perp & v^*H_{12} \\ V_\perp^*H_{11}w & V_\perp^*H_{11}W_\perp & V_\perp^*H_{12} \\ H_{21}w & H_{21}W_\perp & H_{22} \end{bmatrix} \\ = \begin{bmatrix} v^*H_{11}w & v^*H_{11}W_\perp & v^*H_{12} \\ V_\perp^*H_{11}w & \bar{H}_{11} & \bar{H}_{12} \\ H_{21}w & \bar{H}_{21} & H_{22} \end{bmatrix}.$$

One can verify that

$$(30) \quad G_{22}, G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \text{ are all-pass at } \gamma \Rightarrow G_{12} = 0,$$

$$G_{21} = 0 \text{ and } G_{11} \text{ is all-pass at } \gamma.$$

Since V , W , $s_1^{-1}H$, and $s_1^{-1}\bar{H}$ are all-pass, we can use (30) and (29) to show that

$$(31) \quad \begin{bmatrix} V^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} v^*H_{11}w & v^*H_{11}W_\perp & v^*H_{12} \\ V_\perp^*H_{11}w & V_\perp^*H_{11}W_\perp & V_\perp^*H_{12} \\ H_{21}w & H_{21}W_\perp & H_{22} \end{bmatrix} \\ = \begin{bmatrix} s_1 a & 0 & 0 \\ 0 & \bar{H}_{11} & \bar{H}_{12} \\ 0 & \bar{H}_{21} & H_{22} \end{bmatrix},$$

where

$$(32) \quad a := s_1^{-1} v^* H_{11} w \in \mathbb{R} \mathcal{L}_\infty^{1 \times 1} \text{ is all-pass.}$$

Using the characterization of the set $[R_{11} \ R_{12} + \mathbb{S}_1(R)]$ given in Theorem 1, then with a slight abuse of notation, we have

$$(33) \quad \begin{bmatrix} v^* \\ V_\perp^* \end{bmatrix} [R_{11} \ R_{12} + \mathbb{S}_1(R)] \begin{bmatrix} w & W_\perp \end{bmatrix} \\ = \begin{bmatrix} v^* \\ V_\perp^* \end{bmatrix} \{ H_{11} + H_{12} \mathcal{U}_1 (I - H_{22} \mathcal{U}_1)^{-1} H_{21} \} \begin{bmatrix} w & W_\perp \end{bmatrix}.$$

Expanding the right-hand side of (33) and using (31) gives

$$(34) \quad V^* [R_{11} \ R_{12} + \mathbb{S}_1(R)] W = \begin{bmatrix} s_1 a & 0 & 0 \\ 0 & \mathcal{F}_1(\bar{H}, \mathcal{U}_1) \end{bmatrix} = \begin{bmatrix} s_1 a & 0 & 0 \\ 0 & \hat{R}_{11} & \hat{R}_{12} + \mathbb{S}(\hat{R}, s_1) \end{bmatrix},$$

where the second equality follows from part (ii) of Theorem 2. Suppose we write

$$(35) \quad V^* [R_{11} \ R_{12} + Q] W = \begin{bmatrix} s_1 a & 0 & 0 \\ 0 & \hat{R}_{11} & \hat{R}_{12} + \bar{Q} \end{bmatrix}.$$

Then (34) is equivalent to the following. For each $Q \in \mathbb{S}_1(R)$ there exists a $\bar{Q} \in \mathbb{S}(\hat{R}, s_1)$ such that (35) is satisfied. Conversely, for each $\bar{Q} \in \mathbb{S}(\hat{R}, s_1)$, there exists a $Q \in \mathbb{S}_1(R)$ such that (35) is satisfied. Furthermore, as Q ranges over the whole of $\mathbb{S}_1(R)$, \bar{Q} ranges over the whole of $\mathbb{S}(\hat{R}, s_1)$. Since V and W are all-pass, we have

$$s_2(R) := \inf_{Q \in \mathbb{S}_1(R)} s_2^\infty([R_{11} \ R_{12} + Q]) = \inf_{Q \in \mathbb{S}_1(R)} s_2^\infty(V^* [R_{11} \ R_{12} + Q] W).$$

Using (34) and the fact that a is all-pass and fixed,

$$(36) \quad s_2(R) = \inf_{\bar{Q} \in \mathbb{S}(\hat{R}, s_1)} s_1^\infty([\hat{R}_{11} \ \hat{R}_{12} + \bar{Q}]).$$

Since the set $\mathbb{S}(\hat{R}, s_1)$ consists of all suboptimal approximations of \hat{R} , then it includes all the optimal ones. Therefore, the set $\mathbb{S}(\hat{R}, s_1)$ in (36) can be replaced by the set $\mathcal{H}_\infty^{+(p-1) \times (m-1)}$, and (24) follows. Since V and W are all-pass we can write (34) as

$$(37) \quad [R_{11} \ R_{12} + \mathbb{S}_1(R)] = V \begin{bmatrix} s_1 a & 0 & 0 \\ 0 & \hat{R}_{11} & \hat{R}_{12} + \mathbb{S}(\hat{R}, s_1) \end{bmatrix} W^*.$$

Hence, it follows that the subset of $\mathbb{S}_1(R)$ whose elements minimize the second singular value of the left-hand side of (37) can be obtained by replacing the set $\mathbb{S}(\hat{R}, s_1)$ by $\mathbb{S}_1(\hat{R})$ in the right-hand side of (37), which proves that

$$(38) \quad [R_{11} \ R_{12} + \mathbb{S}_2(R)] = V \begin{bmatrix} s_1 a & 0 & 0 \\ 0 & \hat{R}_{11} & \hat{R}_{12} + \mathbb{S}_1(\hat{R}) \end{bmatrix} W^*.$$

To prove (25) we expand the right-hand side of (38) as

$$(39) \quad [R_{11} \ R_{12} + \mathbb{S}_2(R)] = [v \ V_\perp] \begin{bmatrix} s_1 a & 0 & 0 \\ 0 & \hat{R}_{11} & \hat{R}_{12} + \bar{Q}_{12} \end{bmatrix} \begin{bmatrix} w^* \\ W_\perp^* \end{bmatrix} \\ + [v \ V_\perp] \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \mathbb{S}_1(\hat{R}) - \bar{Q}_{12} \end{bmatrix} \begin{bmatrix} w^* \\ W_\perp^* \end{bmatrix}.$$

Using (31), the first term in (39) is equal to $[R_{11} \ R_{12} + Q_{12}]$. Using definition (28) and the partitions of H and \bar{H} , it follows from (39) that

$$[R_{11} \ R_{12} + S_2(R)] = [R_{11} \ R_{12} + Q_{12}] \\ + Q_{13}\bar{Q}_{13}^{-1}[0 \ S_1(\hat{R}) - \bar{Q}_{12}]\begin{bmatrix} \bar{R}_{21}^{-1}R_{21} & * \\ 0 & \bar{Q}_{32}^{-1}Q_{32} \end{bmatrix},$$

where $*$ denotes an expression that is irrelevant for the present purpose. Thus,

$$S_2(R) = Q_{12} + Q_{13}\bar{Q}_{13}^{-1}[S_1(\hat{R}) - \bar{Q}_{12}]\bar{Q}_{32}^{-1}Q_{32} = \mathcal{F}_l[K, S_1(\hat{R})],$$

where

$$K = \begin{bmatrix} Q_{12} - Q_{13}\bar{Q}_{13}^{-1}\bar{Q}_{12}\bar{Q}_{32}^{-1}Q_{32} & Q_{13}\bar{Q}_{13}^{-1} \\ \bar{Q}_{32}^{-1}Q_{32} & 0 \end{bmatrix}.$$

A simple calculation will verify that $\mathcal{F}_l[K, (\cdot)] = \mathcal{F}_l\{Q_a, \mathcal{F}_u[\bar{Q}_a^{-1}, (\cdot)]\}$, which proves (25).

Remark 4. Compared with R , \hat{R} has a lower state dimension $\deg(\hat{R}) \leq \deg(R) - 1$ and

$$R = {}_p[R_{11} \ R_{12}] \in \mathbb{R}\mathcal{H}_{\infty}^{-} \Rightarrow \hat{R} = {}_{p-1}[\hat{R}_{11} \ \hat{R}_{12}] \in \mathbb{R}\mathcal{H}_{\infty}^{-}.$$

The solution of the second-level SODP associated with R is equivalent to the solution of an optimal problem of reduced dimensions and MacMillan degree.

To summarize the results of this section, the following procedure may be used to find the second-level solution of the 2-block SODP. We start with R partitioned as in (4).

(I) Find the optimal level $s_1 = s_1(R)$. If $s_1 = \|R_{11}\|_{\infty}$, stop. If $s_1 > \|R_{11}\|_{\infty}$, continue.

(II) Find H , the generator of all optimal error functions of R using Theorem 1.

(III) Find \bar{H} , the all-pass embedding of H_{22} using Theorem 2 and define $\hat{R} := (\bar{H}_{11})_-$.

The second superoptimal level of R is equal to the optimal level of \hat{R} and the set of all second-level superoptimal approximations of R is given in terms of the optimal approximations of \hat{R} as (25).

Remark 5. The calculation of the optimal levels $s_1(R)$ and $s_1(\hat{R})$ involves a binary search algorithm (the so-called γ -iteration), where each step involves the solution of two algebraic Riccati equations of order $n = \deg(R)$ and $n - 1 = \deg(\hat{R})$, respectively. The calculation of \hat{R} involves the solution of the algebraic Riccati equations (16), which are of order $n - 1$. Hence, the calculation of the second-level superoptimal solution is essentially equivalent to the calculation of the optimal solutions for R and \hat{R} .

Remark 6. If assumption (6) is not satisfied, then the set of all optimal error functions of R can, in general, be generated by an LLFM of the form

$$[R_{11} \ R_{12} + S_1(R)] = \mathcal{F}_l(H, \mathcal{U}), \quad \mathcal{U} := \{[0_{p \times l} \ u]: u \in s_1^{-1} \mathcal{BH}_{\infty}^{p \times m}\},$$

where H is all-pass at s_1 and satisfies [7],

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = {}_p \left[\begin{array}{cc|c} R_{11} & R_{12} + Q_{12} & Q_{13} \\ R_{21} & R_{22} + Q_{22} & Q_{23} \\ \hline 0 & Q_{32} & Q_{33} \end{array} \right] \quad \|H_{22}\|_{\infty} = s_1; H_{12}, H_{21} \text{ have full rank.}$$

It follows that (31) cannot be satisfied for any all-pass transformations V and W since H_{12} and H_{21} have full normal rank. This is because there do not exist vector transfer functions w and v such that the (1, 3) and (3, 1) blocks of (31) are satisfied. Consequently, it is not possible to carry out the diagonalization and the algorithm stops. If $\min(p, m) = 1$, then the optimal and superoptimal approximations are the same and the set of all optimal error functions has a continuum of solutions. This shows that if $s_1(R) = \|R_{11}\|_\infty$, then, in general, there does not exist a unique superoptimal approximation of R . See [11] for more details.

Remark 7. Previous solutions to the SODP are based on finding a maximal Schmidt pair, which are a pair of left and right singular vectors associated with the largest singular value $s_1(R)$ [3], [8], [9], [17]. This pair corresponds to the vectors v and w in (31). Here, in contrast, the solution of the SODP is obtained directly via the optimal generator H ; this is possible since all information about the optimal approximations is contained in this generator (Theorem 1). The advantage of this approach is that the same method can be used for the solution of the one-, two-, and four-block problems, since in each case, the optimal generators have the same form [7].

The solution of the second-level SODP associated with R is given in terms of the solution of the optimal problem associated with \hat{R} . Since \hat{R} has the same form as R , Theorem 1 can be used to give the solution of the optimal problem associated with \hat{R} , and Problem 2 is solved. In § 6, the solution of Problem 2 is used to give the solution of the full SODP (Problem 1).

5. MacMillan degree bounds. In this section, we carry out a cancellation analysis of the second-level superoptimal algorithm developed in the previous section. We will prove the existence of second-level superoptimal approximations of R of MacMillan degree $\leq (n-1) + (n-2)$, where $n = \text{MacMillan degree of } R$. The following lemma gives some properties about linear fractional maps in a state-space setting.

LEMMA 1. Let $P, \hat{P}, \tilde{U} \in \mathbb{R}\mathcal{L}_\infty$ have state-space realizations $\tilde{U} \stackrel{s}{=} (\hat{A}, \hat{B}, \hat{C}, \hat{D})$

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array} \right], \quad \hat{P} = \begin{bmatrix} \hat{P}_{11} & \hat{P}_{12} \\ \hat{P}_{21} & \hat{P}_{22} \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|cc} \hat{A} & \hat{B}_1 & \hat{B}_2 \\ \hline \hat{C}_1 & \hat{D}_{11} & \hat{D}_{12} \\ \hat{C}_2 & \hat{D}_{21} & \hat{D}_{22} \end{array} \right],$$

respectively, and assume that the inverses

$\tilde{L} = (I - D_{22}\tilde{D})^{-1}$, $\tilde{K} = (I - \tilde{D}D_{22})^{-1}$, $\hat{K} = (I - \hat{D}_{11}D_{11})^{-1}$, $\hat{L} = (I - D_{11}\hat{D}_{11})^{-1}$ exist. Then

(i) $\deg(\mathcal{F}_l(P, \tilde{U})) \leq \deg(P) + \deg(\tilde{U})$ and $\deg(\mathcal{F}_u[\hat{P}, \tilde{U}]) \leq \deg(\hat{P}) + \deg(\tilde{U})$.

$$(ii) \mathcal{F}_l(P, \tilde{U}) \stackrel{s}{=} \left[\begin{array}{cc|c} A + B_2\tilde{D}\tilde{L}C_2 & B_2\tilde{K}\tilde{C} & B_1 + B_2\tilde{D}\tilde{L}D_{21} \\ \hline \tilde{B}\tilde{L}C_2 & \tilde{A} + \tilde{B}D_{22}\tilde{K}\tilde{C} & \tilde{B}\tilde{L}D_{21} \\ \hline C_1 + D_{12}\tilde{D}\tilde{L}C_2 & D_{12}\tilde{K}\tilde{C} & D_{11} + D_{12}\tilde{D}\tilde{L}D_{21} \end{array} \right].$$

(iii) $\mathcal{F}_u\{P, \mathcal{F}_l[\hat{P}, (\cdot)]\} = \mathcal{F}_l[J, (\cdot)]$ where

$$J \stackrel{s}{=} \left[\begin{array}{cc|cc} A + B_1\hat{D}_{11}\hat{L}C_1 & B_1\hat{K}\hat{C}_1 & B_2 + B_1\hat{D}_{11}\hat{L}D_{12} & B_1\hat{K}\hat{D}_{12} \\ \hline \hat{B}_1\hat{L}C_1 & \hat{A} + \hat{B}_1D_{11}\hat{K}\hat{C}_1 & \hat{B}_1\hat{L}D_{12} & \hat{B}_2 + \hat{B}_1D_{11}\hat{K}\hat{D}_{12} \\ \hline C_2 + D_{21}\hat{D}_{11}\hat{L}C_1 & D_{21}\hat{K}\hat{C}_1 & D_{22} + D_{21}\hat{D}_{11}\hat{L}D_{12} & D_{21}\hat{K}\hat{D}_{12} \\ \hline \hat{D}_{21}\hat{L}C_1 & \hat{C}_2 + \hat{D}_{21}D_{11}\hat{K}\hat{C}_1 & \hat{D}_{21}\hat{L}D_{12} & \hat{D}_{22} + \hat{D}_{21}D_{11}\hat{K}\hat{D}_{12} \end{array} \right].$$

Proof. The proof follows by routine computation. \square

To give an upper bound on the degree of the second-level superoptimal approximations of R , we need to characterize the set $\mathbb{S}_1(\hat{R})$ as required by Theorem 3. Assume that the optimal level $s_1(\hat{R})$ can be calculated to any desired degree of accuracy using the γ -iteration, and furthermore, that

$$s_2(R) = s_1(\hat{R}) > \|\hat{R}_{11}\|_\infty.$$

Since \hat{R} has the same form as R , Theorem 1 is used to give a characterization of the set $\mathbb{S}_1(\hat{R})$.

COROLLARY 1. *Theorem 1 is satisfied with all variables “hatted”; p , m , and n are replaced by $p - 1$, $m - 1$, and $n - 1$, respectively, and \mathcal{U}_1 and s_1 are replaced by \mathcal{U}_2 and s_2 , respectively.*

The next theorem proves a general cancellation phenomenon between the optimal and suboptimal approximations of a given 2-block system. This result is used to give an upper bound on the MacMillan degree of the second-level superoptimal approximations of R .

THEOREM 4. *Let \bar{Q}_a be the generator of all suboptimal approximations of \hat{R} at level s_1 . Then*

$$(40) \quad \hat{Q} \in \mathbb{S}_1(\hat{R}) \Rightarrow \deg[\mathcal{F}_u(\bar{Q}_a^{-1}, \hat{Q})] \leq \deg(\hat{Q}).$$

Hence,

$$(41) \quad \hat{Q} \in \mathbb{S}_1(\hat{R}) \Rightarrow \deg\{\mathcal{F}_l[Q_a, \mathcal{F}_u(\bar{Q}_a^{-1}, \hat{Q})]\} \leq \deg(\hat{Q}) + n - 1.$$

Finally, if \hat{Q}_{12} is the central optimal approximation of \hat{R} , then

$$(42) \quad \deg\{\mathcal{F}_l[Q_a, \mathcal{F}_u(\bar{Q}_a^{-1}, \hat{Q}_{12})]\} \leq (n - 1) + (n - 2).$$

Proof. Corollary 1 implies that

$$(43) \quad \deg(\hat{Q}_a) \leq n - 2.$$

Since $\hat{Q} \in \mathbb{S}_1(\hat{R})$, then \hat{Q} is generated by the LLFM

$$(44) \quad \hat{Q} = \mathcal{F}_l(\hat{Q}_a, \hat{U}) \text{ for some } \hat{U} \in \mathcal{U}_2.$$

Hence, $\deg(\hat{Q}) \leq \deg(\hat{U}) + n - 2$ from part (i) of Lemma 1. To simplify the proof, assume that, in fact,

$$(45) \quad \deg(\hat{Q}) = \deg(\hat{U}) + n - 2.$$

Using (19), (21), and (1) gives

$$\bar{Q}_a = \begin{bmatrix} \bar{Q}_{12} & \bar{Q}_{13} \\ \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|cc} A_{Q_2} & \bar{B}_{Q_2} & B_{Q_3} \\ \hline \bar{C}_{Q_1} & 0 & s_1 I \\ C_{Q_3} & s_1 I & 0 \end{array} \right] \stackrel{s}{=} \left[\begin{array}{c|cc} -\hat{A}^* - \bar{B}_{Q_2} \hat{B}_2^* & \bar{B}_{Q_2} & B_{Q_3} \\ \hline -\hat{C}_1 \hat{P}_1 & 0 & s_1 I \\ -s_1 \hat{B}_2^* & s_1 I & 0 \end{array} \right]$$

and

$$\bar{Q}_a^{-1} \stackrel{s}{=} \left[\begin{array}{c|cc} -\hat{A}^* + s_1^{-1} B_{Q_3} \hat{C}_1 \hat{P}_1 & s_1^{-1} B_{Q_3} & s_1^{-1} \bar{B}_{Q_2} \\ \hline \hat{B}_2^* & 0 & s_1^{-1} I \\ s_1^{-1} \hat{C}_1 \hat{P}_1 & s_1^{-1} I & 0 \end{array} \right].$$

Part (iii) of Lemma 1, (44), Corollary 1, and (10) yield

$$(46) \quad \mathcal{F}_u(\bar{Q}_a^{-1}, \hat{Q}) = \mathcal{F}_u[\bar{Q}_a^{-1}, \mathcal{F}_l(\hat{Q}_a, \hat{U})] = \mathcal{F}_l(J, \hat{U}),$$

where

$$(47) \quad J \stackrel{s}{=} \left[\begin{array}{cc|cc} s_1^{-1}B_{Q_3}(\hat{C}_1\hat{P}_1 + \hat{D}_{12}\hat{B}_2^*) - \hat{A}^* & s_1^{-1}B_{Q_3}\hat{C}_{Q_1} & s_1^{-1}\bar{B}_{Q_2} + s_1^{-2}B_{Q_3}\hat{D}_{12} & s_1^{-1}B_{Q_3}\hat{D}_{13} \\ \hat{B}_{Q_2}\hat{B}_2^* & \hat{A}_Q & s_1^{-1}\hat{B}_{Q_2} & \hat{B}_{Q_3} \\ \hline s_1^{-1}(\hat{C}_1\hat{P}_1 + \hat{D}_{12}\hat{B}_2^*) & s_1^{-1}\hat{C}_{Q_1} & s_1^{-2}\hat{D}_{12} & s_1^{-1}\hat{D}_{13} \\ \hat{D}_{32}\hat{B}_2^* & \hat{C}_{Q_3} & s_1^{-1}\hat{D}_{32} & 0 \end{array} \right]$$

$$\stackrel{s}{=} \left[\begin{array}{cc|cc} A_{11} & A_{12} & B_{11} & B_{12} \\ A_{21} & A_{22} & B_{21} & B_{22} \\ \hline C_{11} & C_{12} & s_1^{-2}\hat{D}_{12} & s_1^{-1}\hat{D}_{13} \\ C_{21} & C_{22} & s_1^{-1}\hat{D}_{32} & 0 \end{array} \right].$$

Applying the change of basis $T = \begin{bmatrix} I & 0 \\ \hat{P}_3^* & I \end{bmatrix}$ to (47) and using (2),

$$(48) \quad J \stackrel{s}{=} \left[\begin{array}{cc|cc} \hat{A}_{11} & \hat{A}_{12} & \hat{B}_{11} & \hat{B}_{12} \\ \hat{A}_{21} & \hat{A}_{22} & \hat{B}_{21} & \hat{B}_{22} \\ \hline \hat{C}_{11} & \hat{C}_{12} & s_1^{-2}\hat{D}_{12} & s_1^{-1}\hat{D}_{13} \\ \hat{C}_{21} & \hat{C}_{22} & s_1^{-1}\hat{D}_{32} & 0 \end{array} \right]$$

in which

$$(49) \quad \begin{aligned} \hat{A}_{21} &= \hat{A}_Q\hat{P}_3^* + \hat{P}_3^*\hat{A}^* + \hat{B}_{Q_2}\hat{B}_2^* - s_1^{-1}\hat{P}_3^*B_{Q_3}(\hat{C}_1\hat{P}_1 + \hat{D}_{12}\hat{B}_2^* + \hat{C}_{Q_1}\hat{P}_3^*), \\ \hat{C}_{11} &= s_1^{-1}(\hat{C}_1\hat{P}_1 + \hat{D}_{12}\hat{B}_2^* + \hat{C}_{Q_1}\hat{P}_3^*), \quad \hat{C}_{21} = \hat{D}_{32}\hat{B}_2^* + \hat{C}_{Q_3}\hat{P}_3^* \\ \hat{A}_{22} &= \hat{A}_Q - s_1^{-1}\hat{P}_3^*B_{Q_3}\hat{C}_{Q_1}, \quad \hat{C}_{12} = s_1^{-1}\hat{C}_{Q_1}, \end{aligned}$$

$$(50) \quad \begin{aligned} \hat{B}_{21} &= s_1^{-1}[\hat{B}_{Q_2} - \hat{P}_3^*(\bar{B}_{Q_2} + s_1^{-1}B_{Q_3}\hat{D}_{12})], \quad \hat{C}_{22} = \hat{C}_{Q_3}, \\ \hat{B}_{22} &= \hat{B}_{Q_3} - s_1^{-1}\hat{P}_3^*B_{Q_3}\hat{D}_{13}. \end{aligned}$$

It follows from Corollary 1 that the all-pass equations (9) are satisfied for the state-space realization of the generator \hat{H} for some real matrices

$$P_{\hat{H}} = \begin{matrix} n-1 \\ n-2 \end{matrix} \begin{bmatrix} \hat{P}_1 & \hat{P}_3 \\ \hat{P}_3^* & \hat{P}_2 \end{bmatrix}, \quad Q_{\hat{H}} = \begin{matrix} n-1 \\ n-2 \end{matrix} \begin{bmatrix} \hat{Q}_1 & \hat{Q}_3 \\ \hat{Q}_3^* & \hat{Q}_2 \end{bmatrix}.$$

This implies that

$$(51) \quad \begin{aligned} &A_{\hat{H}}P_{\hat{H}} + P_{\hat{H}}A_{\hat{H}}^* + B_{\hat{H}}B_{\hat{H}}^* \\ &= \begin{bmatrix} \hat{A}\hat{P}_1 + \hat{P}_1\hat{A}^* + \hat{B}_1\hat{B}_1^* + \hat{B}_2\hat{B}_2^* & \hat{A}\hat{P}_3 + \hat{P}_3\hat{A}^* + \hat{B}_2\hat{B}_{Q_2} \\ \hat{A}_Q\hat{P}_3^* + \hat{P}_3^*\hat{A}^* + \hat{B}_{Q_2}\hat{B}_2^* & \hat{A}_Q\hat{P}_2 + \hat{P}_2\hat{A}^* + \hat{B}_{Q_2}\hat{B}_{Q_2}^* + \hat{B}_{Q_3}\hat{B}_{Q_3}^* \end{bmatrix} = 0 \end{aligned}$$

and

$$(52) \quad D_{\hat{H}} B_{\hat{H}}^* + C_{\hat{H}} P_{\hat{H}} \\ = \begin{bmatrix} \hat{C}_1 \hat{P}_1 + \hat{C}_{Q_1} \hat{P}_3^* + \hat{D}_{12} \hat{B}_2^* & \hat{C}_1 \hat{P}_3 + \hat{C}_{Q_1} \hat{P}_2 + \hat{D}_{12} \hat{B}_{Q_2}^* + \hat{D}_{13} \hat{B}_{Q_3}^* \\ -s_1^{-1} \hat{B}_1^* \hat{Q}_1 \hat{P}_1 + \hat{C}_{Q_2} \hat{P}_3^* + s_1 \hat{B}_1^* & -s_1^{-1} \hat{B}_1^* \hat{Q}_1 \hat{P}_3 + \hat{C}_{Q_2} \hat{P}_2 \\ \hat{C}_{Q_3} \hat{P}_3^* + \hat{D}_{32} \hat{B}_2^* & \hat{C}_{Q_3} \hat{P}_2 + \hat{D}_{32} \hat{B}_{Q_2}^* \end{bmatrix} = 0.$$

Hence, from (49), the (2, 1) block of (51), and the (1, 1) and (3, 1) blocks of (52), we have $\hat{A}_{21} = 0$, $\hat{C}_{11} = 0$, and $\hat{C}_{21} = 0$, so that (48) and (50) imply that

$$J \stackrel{\Delta}{=} \left[\begin{array}{c|cc} \hat{A}_Q - s_1^{-1} \hat{P}_3^* B_{Q_3} \hat{C}_{Q_1} & s_1^{-1} [\hat{B}_{Q_2} - \hat{P}_3^* (\bar{B}_{Q_2} + s_1^{-1} B_{Q_3} \hat{D}_{12})] & \hat{B}_{Q_3} - s_1^{-1} \hat{P}_3^* B_{Q_3} \hat{D}_{13} \\ \hline s_1^{-1} \hat{C}_{Q_1} & s_1^{-1} \hat{D}_{12} & s_1^{-1} \hat{D}_{13} \\ \hat{C}_{Q_3} & s_1^{-1} \hat{D}_{32} & 0 \end{array} \right].$$

Thus, $\deg(J) \leq n - 2$ since $\hat{A}_Q \in \mathbb{R}^{(n-2) \times (n-2)}$ and this, together with (46), implies that

$$\deg[\mathcal{F}_u(\bar{Q}_a^{-1}, \hat{Q})] \leq \deg(\hat{U}) + n - 2.$$

Hence, (45) proves (40), which, in turn, proves (41) from part (i) of Lemma 1. Finally, (42) follows from (41) since the central approximation $\hat{Q}_{12} \stackrel{\Delta}{=} (\hat{A}_Q, \hat{B}_{Q_2}, \hat{C}_{Q_1}, \hat{D}_{12})$ satisfies $\deg(\hat{Q}_{12}) \leq n - 2$. \square

6. The superoptimal algorithm. In this section, the full 2-block superoptimal algorithm is presented. Let $R_1 = R$, $\hat{R}_1 = \hat{R}$ and $S_1(\cdot) = S(\cdot)$, where $S(\cdot)$ is defined in Theorem 3. The results of §§ 3 and 4 indicate that the solution of the second level SODP associated with R_1 is given in terms of the solution of the ODP associated with \hat{R}_1 as $s_2(R) = s_1(\hat{R})$, $S_2(R) = S_1[S_1(\hat{R})]$, where \hat{R}_1 satisfies

$$\hat{R}_1 = \begin{bmatrix} l & m-1 \\ p-1 & \end{bmatrix} [\hat{R}_{1,11} \quad \hat{R}_{1,12}] \in \mathbb{R}\mathcal{H}_{\infty}^{-}, \quad \deg(\hat{R}_1) \leq n - 1.$$

If we set $R_2 = \hat{R}_1$, we can repeat this process to obtain a sequence of 2-block systems

$$R_i = \begin{bmatrix} l & p-i+1 \\ p-i+1 & \end{bmatrix} [R_{i,11} \quad R_{i,12}] \in \mathbb{R}\mathcal{H}_{\infty}^{-}, \quad i = 1, \dots, k$$

and a sequence of operators S_0, S_1, \dots, S_{k-1} , such that for $j = 1, \dots, k - 1$,

$$(53) \quad s_2(R_j) = s_1(R_{j+1}),$$

$$(54) \quad S_2(R_j) = S_{j-1}[S_1(R_{j+1})],$$

$$(55) \quad \deg(R_j) \leq n - j + 1$$

from Theorems 3 and 4. Hence, a simple induction argument will establish that for $i = 1, \dots, k$,

$$(56) \quad s_i(R_1) = s_1(R_i),$$

$$(57) \quad S_i(R_1) = S_0 \circ S_1 \circ \dots \circ S_{i-1}[S_1(R_i)],$$

where, for operators S_i and S_{i+1} , the operator $S_i \circ S_{i+1}(\cdot)$ is defined by $S_i[S_{i+1}(\cdot)]$. Equation (56) says that the i th superoptimal level for R_1 is equal to the optimal level for R_i , and (57) says that the set of all i th level superoptimal approximations of R_1 can be obtained from the set of all optimal approximations of R_i through the composite operator $S_0 \circ S_1 \circ \dots \circ S_{i-1}$. Thus, we obtain the following algorithm for the solution of the SODP.

ALGORITHM 1. Given $R \in \mathbb{R}\mathcal{H}_\infty^-$ partitioned as in (4):

- (0) Set $i = 1$, $R_1 = R$ and $S_0 = I$;
 (1) Find $s_i = s_1(R_i)$;
 (2) If $i = n$ then set $k = \min(p, m)$; $R_k = R_i$; $s_j = 0$ for $j = n + 1, \dots, k$ and stop;
 (3) If $s_i = \|R_{i,11}\|_\infty$ then set $k = i$ and stop;
 If
 (58) $s_i := s_1(R_i) > \|R_{i,11}\|_\infty$
 then continue;
 (4) If $i = \min(p, m)$ then set $k = \min(p, m)$ and stop;
 (5) Find \hat{R}_i and $S_{i-1}(\cdot)$;
 (6) If it is required to minimize more singular values, then set $R_{i+1} = \hat{R}_i$, $i = i + 1$ and go to step (1); else stop.

The superoptimal levels of R are given by $\{s_1, \dots, s_k\}$ and the set of all k th level superoptimal approximations of R is given by $\mathbb{S}_k(R) = S_0 \circ \dots \circ S_{k-1}[\mathbb{S}_1(R_k)]$.

Next, we generalize the cancellation analysis given in § 5. Theorems 3 and 4, when applied to R_{k-1} , imply that

$$Q \in \mathbb{S}_1(R_k) \Rightarrow S_{k-1}(Q) \in \mathbb{S}_2(R_{k-1}) \text{ and } \deg[S_{k-1}(Q)] \leq \deg(Q) + n - k + 1.$$

Since $\mathbb{S}_2(R_{k-1}) \subseteq \mathbb{S}_1(R_{k-1})$ (a second superoptimal approximation is also an optimal approximation), then by repeated application of Theorems 3 and 4 we obtain

$$(59) \quad Q \in \mathbb{S}_1(R_k) \Rightarrow \deg\{S_0 \circ \dots \circ S_{k-1}[\mathbb{S}_1(R_k)]\} \leq \deg(Q) + \sum_{i=1}^{k-1} n - i.$$

Since $\deg(R_k) \leq n - k + 1$ from (55), then it follows from Theorem 1 that the central optimal approximation $Q_{k,12}$ satisfies $\deg(Q_{k,12}) \leq n - k$. Hence, by taking $Q = Q_{k,12}$ in (59), it follows from (57) that there exist k th level superoptimal approximations of R whose MacMillan degree satisfies

$$(60) \quad \deg(Q_{so}) \leq \sum_{i=1}^k n - i.$$

The number of singular values that can be minimized is given by k , and the superoptimal approximation is unique only if the set $\mathbb{S}_1(R_k)$ has a unique element. If the algorithm stops at the end of step (3), then it follows from Remark 6 that the set $\mathbb{S}_1(R_k)$, in general, has a continuum of solutions, and the superoptimal approximation may not be unique. If the algorithm stops at the end of either steps (2) or (4), then the number of singular values that can be minimized is given by $k = \min(p, m)$, and it follows from Remark 2 that the set $\mathbb{S}_1(R_k)$ has a unique element. Hence, the superoptimal approximation is unique. It also follows from Remark 2 that if the algorithm stops at the end of step (2) then the last $[\min(p, m) - n]$ superoptimal levels are equal to zero.

Remark 8. The algorithm may be stopped prematurely at step (6). In this case, (56), (57), and (60) are still satisfied, which means that there is a saving in the complexity of the algorithm as well as a reduction in the MacMillan degree of the superoptimal approximations. This shows that the algorithm may be used to solve the related problem of minimizing a given number of singular values.

Condition (58) is sufficient for the existence and uniqueness of the superoptimal approximation. To see that this condition is not necessary for existence, consider the system $R = [R_{11} \ R_{12}] = [1/(s-1) \ 0]$. Then it is clear that for this example the

superoptimal approximations are the same as the optimal ones. It can be shown using the results in [7] that $s_1(R) = \|R_{11}\|_\infty = 1$ and that the set of all superoptimal approximations of R is given by $\mathcal{S}_1(R) = \{s/(s+1)u; u \in \mathcal{BH}_\infty^+\}$. Thus, the superoptimal approximation of R exists, although it is not unique. In this case, the superoptimal approach is not appropriate for restoring uniqueness to the optimal approximations, and other criteria must be used. To see that condition (58) is not necessary for uniqueness, consider the system $R = [R_{11} \ R_{12}] = [1/(s-1) \ 1/(s-1)]$. It can be shown using the results in [7] that $s_1(R) = \|R_{11}\|_\infty = 1$ and that the set of all superoptimal approximations of R has a unique element given by $Q_{12} = 1$. Although condition (58) is not satisfied for this example, we have a unique superoptimal approximation. A full discussion is given in [11].

7. The 4-block SODP. This section gives a brief outline of the solution of the 4-block SODP. Given $R \in \mathcal{SH}_\infty^-$ partitioned as

$$R = \begin{array}{c} l \quad m \\ \begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array} \end{array},$$

we are required to find all stable Q that lexicographically minimize the singular values of the error function

$$E_Q = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix}.$$

The optimal level $s_1(R) := \inf_{Q \in \mathcal{H}_\infty^+} s_1^\infty(E_Q)$ is obtained using the γ -iteration [7]. If we assume that

$$s_1 := s_1(R) > \max \{ \| [R_{11} \ R_{12}] \|_\infty, \| [R_{11}^* \ R_{21}^*] \|_\infty \},$$

then [7] gives the set of all optimal error functions of R as the LLFM

$$\mathcal{E}_1(R) := \left\{ E_Q = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix} : Q \in \mathcal{H}_\infty^+, \|E_Q\|_\infty = s_1(R) \right\} = \mathcal{F}_l(H, \mathcal{U}_1),$$

where

$$\mathcal{U}_1 := \left\{ \begin{bmatrix} 0_{q \times l} & 0_{q \times (m-1)} \\ 0_{(p-1) \times l} & u \end{bmatrix} : u \in s_1^{-1} \mathcal{BH}_\infty^{(p-1) \times (m-1)} \right\},$$

and H satisfies $HH^* = H^*H = s_1^2 I$ and has the form

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{array}{c} l \quad m \quad q \quad p-1 \\ \begin{array}{cc|cc} R_{11} & R_{12} & R_{13} & 0 \\ R_{21} & R_{22} + Q_{22} & R_{23} + Q_{23} & Q_{24} \\ \hline R_{31} & R_{32} + Q_{32} & R_{33} + Q_{33} & Q_{34} \\ 0 & Q_{42} & Q_{43} & Q_{44} \end{array} \end{array}.$$

In addition, $Q_{ij} \in \mathcal{RH}_\infty^+$, $i, j = 2, 3, 4$; R_{31}^{-1} , R_{13}^{-1} , $R_{ij} \in \mathcal{RH}_\infty^-$, $i, j = 1, 2, 3$ and $\|H_{22}\|_\infty < s_1$. By using arguments essentially similar to those used in the proof of Theorem 3 we obtain the following result, which gives the solution of the second-level 4-block SODP.

THEOREM 5 [11]. Let H be the generator of all optimal error functions of R . Then

(i) There exists an embedding of H_{22} that is all-pass at s_1 and satisfies

$$\bar{H} = \begin{bmatrix} \bar{H}_{11} & \bar{H}_{12} \\ \bar{H}_{21} & H_{22} \end{bmatrix} = \begin{matrix} & \begin{matrix} l & m-1 & q & p-2 \end{matrix} \\ \begin{matrix} q \\ p-1 \\ l \\ m-2 \end{matrix} & \left[\begin{array}{cc|cc} \hat{R}_{11} & \hat{R}_{12} & \bar{R}_{13} & 0 \\ \hat{R}_{21} & \hat{R}_{22} + \bar{Q}_{22} & \bar{R}_{23} + \bar{Q}_{23} & \bar{Q}_{24} \\ \hline \bar{R}_{31} & \bar{R}_{32} + \bar{Q}_{32} & R_{33} + Q_{33} & Q_{34} \\ 0 & \bar{Q}_{42} & Q_{43} & Q_{44} \end{array} \right] \end{matrix}$$

$\hat{R}_{12}, \hat{R}_{21}, \hat{R}_{22}, \bar{R}_{31}^{-1}, \bar{R}_{13}^{-1}, \bar{R}_{ij} \in \mathbb{RH}_{\infty}^{-}$; $i, j = 1, 2, 3$, $\bar{Q}_{24}^{-1}, \bar{Q}_{42}^{-1}, \bar{Q}_{ij} \in \mathbb{RH}_{\infty}^{+}$; $i, j = 2, 3, 4$, $\hat{R}_{11} \in \mathbb{RL}_{\infty}$.

Hence, the set of all suboptimal error functions of \hat{R} is generated by the LLFM

$$\mathcal{E}(\hat{R}, s_1) := \left\{ \hat{E}_Q = \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \\ \hat{R}_{21} & \hat{R}_{22} + Q \end{bmatrix} : Q \in \mathcal{H}_{\infty}^{+}, \|\hat{E}_Q\|_{\infty} \leq s_1 \right\} = \mathcal{F}_l(\bar{H}, \mathcal{U}_1).$$

(ii) The solution of the second-level SODP associated with R is given by (24) and (25), where

$$Q_a = \begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix}, \quad \bar{Q}_a = \begin{bmatrix} \bar{Q}_{22} & \bar{Q}_{24} \\ \bar{Q}_{42} & \bar{Q}_{44} \end{bmatrix}.$$

Remark 9. A minor complication arises in the solution of the 4-block SODP since \hat{R}_{11} is not necessarily antistable. This can be resolved by noting that

1. There exists an all-pass matrix $X \in \mathbb{RH}_{\infty}^{-}$ such that $\hat{R}_{11} := \hat{R}_{11}X$, $\hat{R}_{21} := \hat{R}_{21}X \in \mathbb{RH}_{\infty}$.

2. The ODP associated with \hat{R} is equivalent to the ODP associated with

$$\tilde{R} := \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \\ \hat{R}_{21} & \hat{R}_{22} \end{bmatrix} = \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \\ \hat{R}_{21} & \hat{R}_{22} \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} \in \mathbb{RH}_{\infty}^{-},$$

since X is all-pass [11].

8. Examples. This section gives a few examples to illustrate the superoptimal algorithm.

Example 1. We give a trivial example as a check, and to illustrate the various steps involved in the superoptimal algorithm. Let

$$R = [R_{11} \quad R_{12}] = \left[\begin{array}{c|c} \frac{4.8}{s-5} & \frac{8}{s-1/2} \\ \hline \frac{-6.4}{s-5} & \frac{6}{s-1/2} \end{array} \middle| \begin{array}{c} \frac{9.6}{s-5} \\ \frac{-12.8}{s-5} \end{array} \right] \in \mathbb{RH}_{\infty}^{-2 \times 3}.$$

Now

$$VR = \begin{bmatrix} .8 & .6 \\ .6 & -.8 \end{bmatrix} R = \left[\begin{array}{c|c} 0 & \frac{1}{s-1/2} \\ \hline \frac{8}{s-5} & \frac{16}{s-5} \end{array} \right],$$

where V is orthogonal. Thus, it follows that $s_i(R) = s_i(VR)$ for $i = 1, 2$. Hence,

$$s_1(R) = s_1(VR) = s_1\left(\left[\frac{8}{s-5} \middle| \frac{16}{s-5}\right]\right) = 2 \quad s_2(R) = s_1(VR) = s_1\left(\frac{1}{s-1/2}\right) = 1.$$

The superoptimal solution can be obtained using the following steps.

1. Using the results in [7], we have that $s_1(R) = 2$ and the generator H is given by

$$H = \left[\begin{array}{cc|c} R_{11} & R_{12} + Q_{12} & Q_{13} \\ R_{21} & R_{22} + Q_{22} & Q_{23} \\ 0 & Q_{32} & Q_{33} \end{array} \right] = \left[\begin{array}{cc|c} \frac{4.8}{s-5} & \frac{.8}{s-1/2} + \frac{4/15}{s+5/6} & \frac{9.6}{s-5} + 1.2 \\ -\frac{6.4}{s-5} & \frac{.6}{s-1/2} + \frac{.2}{s+5/6} & -\frac{12.8}{s-5} - 1.6 \\ 0 & 2\frac{s-3}{s-5} & 0 \end{array} \right] = \left[\begin{array}{cc|c} 2\frac{s-3}{s-5} & 0 & \frac{8}{s-5} \\ 0 & 2\frac{s+1/2}{s+5/6} & 0 \end{array} \right] \begin{array}{c} 1.6\frac{s+1/2}{s+5/6} \\ 1.2\frac{s+1/2}{s+5/6} \\ \frac{4/3}{s+5/6} \end{array}.$$

2. Using Theorem 2, it can be verified that the all-pass embedding of H_{22} (at s_1) is given by

$$\bar{H} = \left[\begin{array}{cc|c} \hat{R}_{11} & \hat{R}_{12} + \bar{Q}_{12} & \bar{Q}_{13} \\ \bar{R}_{21} & \bar{R}_{22} + \bar{Q}_{22} & \bar{Q}_{23} \\ 0 & \bar{Q}_{32} & \bar{Q}_{33} \end{array} \right] = \left[\begin{array}{cc|c} 0 & \frac{1}{s-1/2} + \frac{1/3}{s+5/6} & 2\frac{s+1/2}{s+5/6} \\ 2 & 0 & 0 \\ 0 & 2\frac{s+1/2}{s+5/6} & \frac{4/3}{s+5/6} \end{array} \right].$$

3. The second-level system \hat{R} is given by the antistable projection of \bar{H}_{11} so that

$$\hat{R} = [\bar{H}_{11}]_- = \left[\begin{array}{c} 1 \\ s - 1/2 \end{array} \right],$$

which defines an effectively one-block problem since $\hat{R}_{11} = 0$. Then it follows that

$$s_1(\hat{R}) = s_1\left(\frac{1}{s - 1/2}\right) = 1,$$

and it can be checked using Corollary 1 that the set $\mathbb{S}_1(\hat{R})$ has only one element given by $\hat{Q}_{12} = 1$.

4. Finally, we use Theorem 3 to obtain the second superoptimal level $s_2(R)$ as $s_2(R) = s_1(\hat{R}) = 1$ and the (unique) second-level superoptimal approximation of R as

$$S_2(R) = Q_{12} + Q_{13}\bar{Q}_{13}^{-1}[\hat{Q}_{12} - \bar{Q}_{12}]\bar{Q}_{32}^{-1}Q_{32} = \begin{bmatrix} .8 & 1.2 \\ .6 & -1.6 \end{bmatrix}.$$

Example 2. Let $R \in \mathbb{R}\mathcal{H}_\infty^{-2 \times 3}$ be given by

$$R = [R_{11} \quad R_{12}] = \left[\begin{array}{cc} \frac{2.5}{s-1.8} + \frac{0.5}{s-3.4} & \frac{12.5}{s-1.8} + \frac{0.5}{s-3.4} \\ \frac{0.5}{s-1.8} + \frac{3.5}{s-3.4} & \frac{2.5}{s-1.8} + \frac{3.5}{s-3.4} \end{array} \right] \begin{array}{c} \frac{2.5}{s-1.8} + \frac{3.5}{s-3.4} \\ \frac{0.5}{s-1.8} + \frac{24.5}{s-3.4} \end{array}.$$

Then by carrying out the steps given in this paper, the superoptimal levels are given by $s_1(R) = 5.0$, $s_2(R) = 2.40157$, and the (unique) superoptimal approximation of R is obtained as

$$Q_{so} = \frac{1}{(s + 2.46868)} \begin{bmatrix} 3.33701(s + 2.53702) & 1.24724(s + 2.66313) \\ 1.24724(s + 2.66313) & 4.06457(s + 2.53215) \end{bmatrix}.$$

A plot of the singular values of the superoptimal error system $[R_{11} \ R_{12} + Q_{so}]$ will verify that they are flat at the superoptimal levels. \square

9. Conclusion. Here we summarize the main contributions of this paper.

A new and simple algorithm is presented for the solution of the 2-block superoptimal distance problem. The algorithm reduces the superoptimal problem into a series of optimal problems, one for each extra singular value to be minimized. It is shown that the superoptimal solution is unique if each of these optimal problems has an optimal level satisfying (58).

An expression is obtained for the upper bound on the MacMillan degree of the superoptimal approximation. This gives a simple generalization of the bound already obtained for the 1-block problem [8], [15].

The superoptimal algorithm is shown to be truly recursive in that it can be stopped after minimizing a given number of singular values. In this case, there is a predictable saving in the computation and a corresponding reduction in the MacMillan degree of the approximation.

It is shown that the algorithm can be extended readily to the solution of the 4-block SODP.

Apart from providing a way of restoring uniqueness to the optimal approximations, superoptimal solutions can offer advantages over the optimal ones. Some work has already been done regarding the application of superoptimal approximations in control design problems. For example, in [17] superoptimality is justified for problems involving multiobjective disturbance rejection. In [9], the results in [12] are used to relate superoptimal approximations to robust stability \mathcal{H}_∞ design. The authors are currently considering the potential applications of the superoptimal approximations to model reduction problems.

REFERENCES

- [1] B. D. O. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis. A Modern Systems Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [2] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard \mathcal{H}^2 and \mathcal{H}^∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [3] Y. K. FOO AND I. POSTLETHWAITE, *An \mathcal{H}^∞ -minimax approach to the design of robust control systems*, Systems Control Lett., 5 (1984), pp. 81–82.
- [4] B. A. FRANCIS, *A Course in \mathcal{H}_∞ Control Theory*, Springer-Verlag, Berlin, 1987.
- [5] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their \mathcal{L}^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.
- [6] ———, *Robust stabilization of linear multivariable systems: Relations to approximation*, Internat. J. Control, 43 (1986), pp. 741–766.
- [7] K. GLOVER, D. J. N. LIMEBEER, J. C. DOYLE, E. M. KASSENALLY, AND D. J. SAFONOV, *A characterization of all solutions to the four-block general-distance problem*, SIAM J. Control Optim., 29 (1991), pp. 283–324.
- [8] D. W. GU, M. C. TSAI, AND I. POSTLETHWAITE, *An algorithm for super-optimal \mathcal{H}_∞ design: The 2-block case*, Automatica, 26 (1990), pp. 437–440.
- [9] G. D. HALIKIAS, *State-space analysis for a class of \mathcal{H}_∞ optimal control problems*, Ph.D. thesis, Imperial College, London University, 1990.
- [10] Y. S. HUNG, *\mathcal{H}^∞ interpolation of rational matrices*, Internat. J. Control, 48 (1988), pp. 1659–1713.
- [11] I. M. JAIMOUKHA, *The two-block super-optimal distance problem*, Ph.D. thesis, Imperial College, London University, 1990.
- [12] N. LEHTOMAKI, N. SANDEL, AND M. ATHANS, *Robustness results in linear-quadratic Gaussian based multivariable control designs*, IEEE Trans. Automat. Control, 26 (1981), pp. 75–93.
- [13] D. J. N. LIMEBEER, E. M. KASSENALLY, I. JAIMOUKHA, AND M. G. SAFONOV, *A characterization of all*

- solutions to the four block general distance problem*, Proc. IEEE Conference on Decision and Control, Austin, TX, 1988.
- [14] D. J. N. LIMEBEER AND G. D. HALIKIAS, *An analysis of pole-zero cancellations in \mathcal{H}_∞ -optimal control problems of the second kind*, SIAM J. Control Optim. 26 (1988), pp. 646–677.
 - [15] D. J. N. LIMEBEER, G. D. HALIKIAS, AND K. GLOVER, *State-space algorithm for the computation of super-optimal matrix interpolating functions*, Internat. J. Control, 50 (1989), pp. 2431–2466.
 - [16] Z. NEHARI, *On bounded bilinear forms*, Ann. of Math., 65 (1984), pp. 153–162.
 - [17] M. C. TSAI, D. W. GU, AND I. POSTLETHWAITE, *A state space approach to super-optimal H^∞ control problems*, IEEE Trans. Automat. Control 33 (1988), pp. 833–843.
 - [18] D. C. YOULA, H. JABR, AND J. J. BONGIORNO, *Modern Wiener-Hopf design of optimal controllers, Part II: The multivariable case*, IEEE Trans. Automat. Control, 21 (1976), pp. 319–338.
 - [19] N. J. YOUNG, *The Nevanlinna-Pick problem for matrix-valued functions*, J. Operator Theory, 15 (1986), pp. 239–265.

PROPERTIES OF RELAXED TRAJECTORIES OF EVOLUTION EQUATIONS AND OPTIMAL CONTROL*

X. XIANG† AND N. U. AHMED‡

Abstract. This paper presents a density result for a general class of nonlinear infinite-dimensional systems that includes an unbounded linear part. A result on the existence of optimal controls for a corresponding uncertain system is also proved.

Key words. nonlinear, infinite-dimensional systems, relaxed trajectories, denseness, existence, optimal controls, uncertain systems

AMS subject classifications. 93C25, 49J20, 49J27

1. Introduction. Recently, in a very interesting paper, Papageorgiou [8] has studied the properties of relaxed trajectories of a general class of evolution equations in optimal control. He detected an error in an earlier paper of Ahmed [2] on the same topic and succeeded in giving a correct proof following the same procedure as Ahmed. In the process, however, Papageorgiou introduced several stronger assumptions on both the operators A and f (see [8, p. 271]), including compactness of the embedding $X \hookrightarrow H$. Hence the system considered by Papageorgiou does not cover the original model considered by Ahmed [2]. Later in his paper, Papageorgiou extends his results to a more general class $(**)$ (see [8, p. 279]), where he allows a fully nonlinear system rather than the semilinear one $(*)$ (see [8, p. 271]). Here, in the general case, A is assumed to have the same regularity properties (see [8, p. 279, $H(A_1)$]) as those used by Ahmed for the nonlinear part (see Ahmed [2]). In other words, the system considered by Ahmed has an additional term representing a linear unbounded operator that is missing from the fully nonlinear model of Papageorgiou. The nonlinear operator in both Papageorgiou and Ahmed is, in fact, bounded. This means that Papageorgiou's model is covered by the model considered in Ahmed's paper.

In this paper we present a density result for the more general system that includes the unbounded linear part. Furthermore, we also present a result on the existence of optimal controls for a class of uncertain systems.

2. Preliminaries. Let H be a separable Hilbert space and E a dense linear subspace of H carrying the structure of a reflexive Banach space with H^* and E^* denoting the corresponding (topological) duals. Identifying H with its dual, we have $E \hookrightarrow H \hookrightarrow E^*$ with the embeddings being continuous and dense. By $\langle \cdot, \cdot \rangle$ we will denote the duality brackets for the pair (E, E^*) and by (\cdot, \cdot) the inner product in H . The two are compatible in the sense that $\langle \cdot, \cdot \rangle_{E, H} = (\cdot, \cdot)_H$. The norms will be denoted by $\|\cdot\|_G$ for $G = (E, H, E^*)$. The Banach space G furnished with the weak topology will be denoted by G_w .

Let B be a Polish space and Γ a closed subset of B . Let U denote the set of all measurable functions u defined on $I = (0, T) \subset \mathbb{R}^+ = [0, \infty]$ with values $u(t) \in \Gamma$ almost everywhere. We call U the class of original controls. In this paper we assume Γ is compact. An example of a compact Polish space is the unit ball $\Gamma \equiv B_1(Y)$ of a Banach space Y

* Received by the editors October 7, 1991; accepted for publication (in revised form) April 8, 1992. This work was supported in part by National Science and Engineering Research Council of Canada grant 7109.

† Department of Mathematics, Guizhou University, Guiyang, Guizhou, People's Republic of China.

‡ Department of Electrical Engineering, Department of Mathematics and Department of Systems Science, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5.

having a separable dual Y^* and furnished with the weak topology. For many other examples see [2].

Let $M(\Gamma)$ denote the space of all probability measures on the Borel σ -field of Γ and $C(\Gamma)$ the space of all real-valued bounded continuous functions on Γ . Suppose $M(\Gamma)$ is furnished with the w^* -topology. This topology is metrizable and $M(\Gamma)$, with this topology, becomes a compact Polish space. Let \mathcal{M} denote the family of all weakly measurable functions $\{\mu_t\}$ defined on I with values $\mu_t \in M(\Gamma)$ for all $t \in I$ (see [2]).

Let $B(I \times \Gamma)$ denote the (topological) Borel algebra of subsets of the set $I \times \Gamma$ and consider the measurable space $(I \times \Gamma, B(I \times \Gamma))$ and let $R(I, \Gamma)$ denote the space of all transition measures $\{\mu_t(d\sigma)dt\}$ on $B(I \times \Gamma)$ satisfying the following properties:

- (1) $\mu_t(\Gamma) = 1$ for a.e. $t \in I$;
- (2) $\int_{I \times \Gamma} \chi_{J \times \Gamma}(t, \sigma) \mu_t(d\sigma) dt = l(J)$, $J \in B(I)$;
- (3) $\int_{I \times \Gamma} g(t, \sigma) \mu_t(d\sigma) dt \geq 0$ for all $g \geq 0$,

where l denotes the Lebesgue measure and $\chi_{J \times \Gamma}$ the characteristic function of the set $J \times \Gamma$. The space $R(I, \Gamma)$ is furnished with the weakest topology, making the functionals $I_f(\mu) = \int_0^T \int_{\Gamma} f(t, \sigma) \mu_t(d\sigma) dt$ (for any Caratheodory integrand f , that is, functions measurable in t on I and continuous in σ on Γ) continuous (see [3]).

Since $\bar{I} = [0, T] \subseteq \mathbb{R}^+$ and Γ is a compact Polish space, the Caratheodory integrands on $I \times \Gamma$ may be identified with Lebesgue–Bochner space $L^1(I, C(\Gamma))$, abbreviated as $L^1(C(\Gamma))$. Now, by the Riesz representation theorem, $[C(\Gamma)]^* = M(\Gamma)$, the space of all bounded Borel measures on $B(\Gamma)$. $M(\Gamma)$ has the Radon–Nikodym property. Thus $(L^1(C(\Gamma)))^* = L^\infty(M(\Gamma)) \equiv L^\infty(I, M(\Gamma))$. So the weak topology on $R(I, \Gamma)$ coincides with relative topology $\sigma(L^\infty(M(\Gamma)), L^1(C(\Gamma)))$ (see [4] and [5]).

We consider the controlled system governed by the following nonlinear evolution equation:

$$(1) \quad dx/dt = A(t)x + f(t, x, u), \quad t \in I, \quad x(0) = x_0,$$

in the Banach space E , where $u \in U$, $\{A(t), t \in I\}$ is a family of densely defined linear operators with domain $D(A(t)) \subset E$ and range $R(A(t)) \subset E^*$ and, in general, $f: I \times E \times \Gamma \rightarrow E^*$.

Similarly, for a $\mu \in \mathcal{M}$, we consider the following evolution equation:

$$(2) \quad dx/dt = A(t)x + \int_{\Gamma} f(t, x, \sigma) \mu_t(d\sigma), \quad t \in I, \quad x(0) = x_0.$$

We call this the relaxed system.

For $1 < p, q < \infty$ with $1/p + 1/q = 1$, let L denote the operator determined by

$$D(L) = \{x \in L_p(I, E), x(t) \in D(A(t)) \cap D(A^*(t)),$$

$$\text{for a.e. } t \in I \text{ and } A(t)x(t), A^*(t)x(t), \dot{x}(t) \in C^0(I, E^*) \cap L_q(I, E^*)\}$$

with $(Lx)(t) = (d/dt - A(t))x, t \in I$ for $x \in D(L)$.

We assume throughout the paper that L is densely defined as a linear operator from $L_p(I, E)$ to $L_q(I, E^*)$ and that the strong and weak closures (sometimes called extensions) of L from $L_p(I, E)$ to $L_q(I, E^*)$ coincide (that is, $L_s = L_w$).

For a discussion on this assumption see the remark below.

For $x_0 \in D(A(t))$, $t \in I$, $u \in U$, an element $x_u \in L_p(I, E)$ is said to be a strong solution of the evolution equation (1) if $x_u(0) = x_0$ and $(L_s x_u)(t) = f(t, x_u(t), u(t))$ almost everywhere on I . We denote this family of solutions by $X = \{x_u | u \in U\}$ and call it the set of original trajectories. Similarly, for a generalized control $\mu \in \mathcal{M}$, an element $x_\mu \in L_p(I, E)$ is a strong solution of the evolution equation (2) if $x_\mu(0) = x_0$ and

$(L_s x_\mu)(t) = \int_\Gamma f(t, x_\mu(t), \sigma) \mu_t(d\sigma)$ almost everywhere on I . We denote this family of solutions by $X_r = \{x_\mu | \mu \in \mathcal{M}\}$ and call it the set of relaxed trajectories.

3. Existence and properties of solutions. We assume throughout the presentation, unless stated otherwise, that the operators A and f satisfy the following conditions.

CONDITION A1. $\{A(t), t \in I\}$ is a family of densely defined linear operators in H (not necessarily bounded) with domain $D(A(t)) \subset E$ and range $R(A(t)) \subset E^*$ for $t \in I$.

CONDITION A2. $\langle A(t)e, e \rangle_{E^*, E} \leq 0$ for all $e \in D(A(t)) \subset E$, $t \in I$ and the strong and weak extensions or equivalently closures of L from $L_p(I, E)$ to $L_q(I, E^*)$ coincide.

CONDITION F1. The function $t \rightarrow \langle f(t, e, \sigma), g \rangle_{E^*, E}$ is continuous on I for arbitrary $e, g \in E$ and $\sigma \in B$ and $f: I \times E \times B \rightarrow E^*$ is demicontinuous in the sense that whenever $t_n \rightarrow t$ in I , $\xi_n \rightarrow \xi$ in E and $v_n \rightarrow v$ in B

$$\langle f(t_n, \xi_n, v_n), e \rangle_{E^*, E} \rightarrow \langle f(t, \xi, v), e \rangle_{E^*, E}$$

for each $e \in E$.

CONDITION F2.

$$\langle f(t, x, \sigma) - f(t, y, \sigma), x - y \rangle_{E^*, E} \leq 0$$

for all $x, y \in E$ and $\sigma \in \Gamma$.

CONDITION F3. There exists an $h \in \Gamma_q(I, R^+)$ and $\alpha \geq 0$ such that

$$\|f(t, x, \sigma)\|_{E^*} \leq h(t) + \alpha \|x\|_E^{p/q}$$

for each $x \in E$ and for all $\sigma \in \Gamma$.

CONDITION F4. There exists an $h_1 \in L_1(I, R^+)$, $\beta > 0$, such that

$$\langle f(t, x, \sigma), x \rangle_{E^*, E} \leq h_1(t) - \beta \|x\|_E^p \text{ a.e.}$$

for each $x \in E$ and for all $\sigma \in \Gamma$.

CONDITION F5. If $e_n \rightarrow e$ in E_w , then, for each $x \in E$ and almost all $t \in I$,

$$\langle f(t, x, \sigma), e_n \rangle_{E^*, E} \rightarrow \langle f(t, x, \sigma), e \rangle_{E^*, E}$$

uniformly with respect $\sigma \in \Gamma$.

Remark on the identity $L_s = L_w$. This assumption was introduced by Browder in the study of a very general class of nonlinear evolution equations on Banach spaces (see Browder [10]–[12]). The proof of this result is by no means trivial. Several conditions under which this equivalence is valid are given in Browder [10, Thms. 6 and 7, pp. 73–77]. In an earlier paper (see Browder [11]), the evolution equation was considered in the Hilbert space setting, and in this case the coincidence of the weak and strong extensions was also proved (see [11, Thm. 5, p. 508]). For a discussion on the same topic see also Browder [12, Thm. 7, p. 39]. Browder's result was also extended by Kato [13], where $A(t)$ for each $t \in I$ was assumed to be the infinitesimal generator of a contraction semigroup along with the hypothesis that $(\lambda I - A(t))^{-1}$ is C^1 in $t \in I$ for $\lambda > 0$ in the strong operator topology. For a complete proof of this result see the proof of Theorem 4 of Kato [13, Thm. 4, pp. 61–65] where, among other things, the equivalence is proved.

LEMMA 3.1. Under Conditions A1, A2, and F1–F4,

- (1) for each (original) control $u \in U$ and initial state $x_0 \in E$, the evolution equation (1) has a unique strong solution $x_u \in L_p(I, E)$;
- (2) for each (relaxed) control $\mu \in \mathcal{M}$, and initial state $x_0 \in E$, the evolution equation (2) has a unique strong solution $x_\mu \in L_p(I, E)$.

Furthermore, in either case, $x_u, x_\mu \in C(I, E_w) \cap C(\bar{I}, H)$ and $x_u, x_\mu \in D(A)$ almost everywhere (see [2, Lemma 4.1]).

LEMMA 3.2. Suppose the operators A and f satisfy Conditions A1, A2, and F1–F4, respectively. Then the set of original trajectories X and the set of relaxed trajectories X_r are all conditionally sequentially compact subsets of $C(I, E_w)$. Furthermore, X_r is a sequentially compact subset of $C(I, E_w)$. (See Lemma 5.1 and Theorem 5.1 of [2].)

Now we give some properties of relaxed trajectories that are useful in the study of optimal controls.

THEOREM 3.1. Under Conditions A1, A2, and F1–F5 the family of original (relaxed) trajectories X (X_r) are bounded in $L_p(I, E)$ and $C(\bar{I}, H)$. Furthermore, whenever $\mu^n \xrightarrow{w} \mu^0$ in $R(I, \Gamma)$, the sequence x_{μ^n} has a subsequence converging to x_{μ^0} in the topology of $C(\bar{I}, H)$.

Proof. Let $x \in X$. It follows from (1) and Conditions A2 and F4, that

$$\begin{aligned} d/dt \|x(t)\|_H^2 &= 2(\langle A(t)x(t), x(t) \rangle + \langle f(t, x(t), u(t)), x(t) \rangle) \\ &\leq 2\langle f(t, x(t), u(t)), x(t) \rangle \leq 2(h_1(t) - \beta \|x(t)\|_E^p) \end{aligned}$$

for $0 \leq t \leq T$. Hence

$$\|x(t)\|_H^2 + 2\beta \int_0^t \|x(s)\|_E^p ds \leq \|x(0)\|_H^2 + 2 \int_0^t h_1(s) ds.$$

This inequality shows that X is a bounded subset of $L_\infty(\bar{I}, H) \cap L_p(I, E)$.

Furthermore, since $Ax \in L_q(I, E^*)$ for $x \in X$, we have $\{\dot{x}\} \in L_q(I, E^*)$ for $x \in X$. Hence $x \in C(\bar{I}, H)$ and X is a bounded subset of $C(\bar{I}, H)$. Similarly, X_r is a bounded subset of $L_p(I, E) \cap C(\bar{I}, H)$. For abbreviation, let $x_n = x_{\mu^n}$, $x_0 = x_{\mu^0}$. Using Conditions A2 and F2, we can verify that

$$\begin{aligned} &d/dt (\|x_0(t) - x_n(t)\|_H^2) \\ (*) \quad &\leq 2 \int_\Gamma \langle f(t, x_0, \sigma), x_0 - x_n \rangle \mu_t^0(d\sigma) - 2 \int_\Gamma \langle f(t, x_0, \sigma), x_0 - x_n \rangle \mu_t^n(d\sigma). \end{aligned}$$

By Lemma 3.2 we may assume that $x_n \rightarrow x^*$ in $C(I, E_w)$. Define

$$g_n(t, \sigma) = \langle f(t, x_0(t), \sigma), x_0(t) - x_n(t) \rangle$$

and

$$g(t, \sigma) = \langle f(t, x_0(t), \sigma), x_0(t) - x^*(t) \rangle.$$

By passing to a subsequence if necessary, it follows from Condition F5 that

$$g_n(t, \sigma) \rightarrow g(t, \sigma) \quad \text{in } C(\Gamma)$$

for almost all $t \in I$.

Consider the family of operators $\{T_n, n \in N \cup \{0\}\}$ from $\mathcal{L}(E^*, C(I))$ defined by

$$(T_n e^*)(t) = \langle x_n(t), e^* \rangle, \quad t \in I, \quad e^* \in E^*.$$

Since, by Lemma 2.2, X_r is a sequentially compact subset of $C(I, E_w)$, there exist constants k_{e^*} independent of n such that

$$\|T_n e^*\|_{C(I)} = \sup \{ |\langle x_n(t), e^* \rangle|, t \in I \} \leq k_{e^*}.$$

Then it follows from the uniform boundedness principle (see [1, Thm. 1.1.3]) that for all $n \in N \cup \{0\}$

$$\sup \|x_n(t)\|_E \leq K$$

for some constant K independent of n and t .

Hence, combining Lemma 2.2 and Condition F3, we have

$$\begin{aligned} |g_n(t, \sigma)| &= |\langle f(t, x_0(t), \sigma), x_0 - x_n \rangle| \\ &\leq 2K[h(t) + \alpha \|x_0(t)\|_E^{p/q}] = G(t). \end{aligned}$$

Obviously, $G \in L^1$. Hence, by virtue of the dominated convergence theorem, we obtain

$$(**) \quad g_n(t, \sigma) \rightarrow g(t, \sigma) \quad \text{in } L^1(C(\Gamma)).$$

Since $\mu^n \xrightarrow{w} \mu^0$ in $R(I, \Gamma)$, $\mu^n \xrightarrow{w^*} \mu^0$ in $L^\infty(M(\Gamma)) = [L^1(C(\Gamma))]^*$, for each $t \in I$,

$$(***) \quad \int_0^t \int_\Gamma g_n(s, \sigma) \mu_s^n(d\sigma) ds \rightarrow \int_0^t \int_\Gamma g(s, \sigma) \mu_s^0(d\sigma) ds.$$

Combining (*)–(***) , we have

$$x_n \rightarrow x_0 \quad \text{in } C(\bar{I}, H).$$

This completes the proof.

Notes.

1. If $\sigma \rightarrow f(t, x, \sigma)$ is continuous, then Condition F5 holds.
2. If $f: I \times E \times \Gamma \rightarrow H$ and the injection $E \hookrightarrow H$ is compact, then the same conclusion holds.

This last assumption was used in [8].

COROLLARY 3.1. *If the assumptions of Theorem 3.1 hold, then X_r is sequentially compact in $C(\bar{I}, H)$.*

Proof. From Theorem v-1 of Castaing and Valadier [6] we know that \mathcal{M} is w^* -compact in $L^\infty(M(\Gamma))$ and w -compact in $R(I, \Gamma)$. Using Theorem 3.1 we immediately reach the conclusion.

Finally, we give a denseness result for relaxed trajectories.

THEOREM 3.2. *Suppose Conditions A1, A2, and F1–F5 hold; then X is dense in X_r with respect to the usual topology of $C(\bar{I}, H)$.*

Proof. Let $x_0 \in X_r$, that is $x_0 = x_{\mu^0}$ for some $\mu^0 \in R(I, \Gamma) = L^\infty(M(\Gamma))$. By virtue of a result of Balder [7, Cor. 3], we can find $u_n \in U$ such that $\delta_{u_n} \xrightarrow{w^*} \mu^0$ in $R(I, \Gamma)$. Letting $y_n = x_{u_n}$, we have

$$\dot{x}_0(t) = A(t)x_0(t) + \int_\Gamma f(t, x_0(t), \sigma) \mu_t^0(d\sigma),$$

and

$$\begin{aligned} \dot{y}_n(t) &= A(t)y_n(t) + f(t, y_n(t), u_n) \\ &= A(t)y_n(t) + \int_\Gamma f(t, y_n(t), \sigma) \delta_{u_n}(d\sigma). \end{aligned}$$

Using the same procedure as in Theorem 3.1, and by passing to a subsequence if necessary, we can verify that

$$y_n \rightarrow x_0 \quad \text{in } C(I, H).$$

This means that X is dense in X_r , proving the assertion.

4. Existence of relaxed controls (uncertain system). In modeling physical systems, simplifying assumptions often are made to obtain a manageable differential or integral or a functional equation. Thus the behavior of the model may not precisely coincide with the experimental observations. Even if the exact form and order of the differential

equation is known, often the exact parameter values may be unknown due to the unavailability of basic scientific data. For example, in the case of a nonlinear heat equation, the conductivity of the material may be a function of temperature itself and its exact form may be unknown. Similarly, the modulus of rigidity of a beam made of trusses or nonhomogeneous materials is very difficult to determine experimentally. This may be further complicated by the presence of loose bolts and nuts or defective welding. In the Schrödinger equation, the precise functional form of the nuclear or Coulomb potential for many particle systems may be unknown. These facts lead to the concept of "uncertain" systems.

An uncertain version of the relaxed control system (2) may be given by

$$(3) \quad dx/dt = A(t)x + \int_{\Gamma} f(t, x, \sigma, z) \mu_t(d\sigma), \quad t \in I, \quad x(0) = x_0, \quad z \in \mathcal{Z}.$$

The set \mathcal{Z} may be another compact Polish space or, more generally, a Souslin space. Recall that a Hausdorff topological space \mathcal{Z} is a Souslin space if there exists a Polish space Y and a continuous surjection of Y to \mathcal{Z} . A Souslin space is always separable, an example being a separable Banach space endowed with the weak topology. The uncertainty of the system is reflected in the values the parameter z of f may take from the set \mathcal{Z} . A fact that distinguishes an uncertain system from that of a stochastic system is that even the probability measure on $B(\mathcal{Z})$ that could describe the uncertainty is not known. In fact, an uncertain system may be better described by a differential inclusion [9].

Consider the relaxed control system (3) and suppose that A satisfies Conditions A1 and A2 and $f(\cdot, \cdot, \cdot, z)$, $z \in \mathcal{Z}$, satisfies Conditions F1–F5 uniformly with respect to $z \in \mathcal{Z}$, with h, h_1, α, β possibly depending on the set \mathcal{Z} but not on its individual elements. Clearly, under these conditions the results of the preceding sections hold for each $z \in \mathcal{Z}$. Let x_{μ}^z denote the solution of the relaxed system (3) corresponding to $z \in \mathcal{Z}$ and $\mu \in \mathcal{M}$. Let $\mathcal{X}_{\mu} \equiv \{x_{\mu}^z, z \in \mathcal{Z}\}$ denote the family of solutions of the uncertain system (3) corresponding to a fixed relaxed control $\mu \in R(I, \Gamma)$. Let $l: I \times H \times \Gamma \rightarrow R$ be a Borel measurable map. Define the functional

$$J(x, \mu) = \int_I \int_{\Gamma} l(t, x, \sigma) \mu_t(d\sigma) dt$$

corresponding to $\mu \in \mathcal{M}$ and $x \in \mathcal{X}_{\mu}$. Since, a priori, it is not known which element z of \mathcal{Z} is in force, all that a system analyst can do is minimize the maximum risk. Thus the objective functional is taken as

$$J_0(\mu) \equiv \text{Sup} \{J(x, \mu), x \in \mathcal{X}_{\mu}\}.$$

We consider the question of existence of an optimal relaxed control $\mu^0 \in \mathcal{M}$ in the sense that

$$J_0(\mu^0) \leq J_0(\mu) \quad \text{for all } \mu \in \mathcal{M}.$$

We will make the following hypotheses concerning the integrand $l(\cdot, \cdot, \cdot)$.

CONDITION L. $l: I \times H \times \Gamma \rightarrow R$ maps bounded subsets of $I \times H \times \Gamma$ into bounded subsets of R and satisfies the following assumptions:

(1) $(t, x, \sigma) \rightarrow l(t, x, \sigma)$ is measurable with respect to the Borel field $B(I) \times B(H) \times B(\Gamma)$;

(2) $(x, \sigma) \rightarrow l(\cdot, x, \sigma)$ is lower semicontinuous;

(3) $\psi(t) \leq l(t, x, \sigma)$ almost everywhere with $\psi(\cdot) \in L^1$.

Since $f(\cdot, \cdot, \cdot, z)$, $z \in \mathcal{Z}$ satisfies Conditions F1–F5 uniformly with respect to the set \mathcal{Z} , for each $\mu \in \mathcal{M}$, the set \mathcal{X}_{μ} is a bounded subset of $C(I, H) \cap L_p(I, E)$. Hence it

follows from Theorem 3.1 and Condition L that the cost functional $J_0(\mu)$ is well defined and it is bounded for all $\mu \in \mathcal{M}$. Define

$$\inf \{J_0(\mu), \mu \in \mathcal{M}\} \equiv m.$$

THEOREM 4.1. *Suppose the operators A, f, l satisfy Conditions A1, A2, F1–F5, and L, respectively. Then there exists a $\mu^0 \in \mathcal{M}$ such that $J_0(\mu^0) = m$.*

Proof. Let $\{\mu^n\}$ be a minimizing sequence for the functional $J_0(\mu)$, that is,

$$\lim_{n \rightarrow \infty} J_0(\mu^n) = \inf \{J_0(\mu), \mu \in \mathcal{M}\} = m.$$

Recall that \mathcal{M} is w^* -compact in $L^\infty(M(\Gamma))$ and w -compact in $R(I, \Gamma)$. By passing to a subsequence if necessary, we may assume $\mu^n \xrightarrow{w} \mu^0$ in $R(I, \Gamma)$.

For any $\varepsilon > 0$, there exists $x_0 \in \mathcal{X}_{\mu^0}$ and hence $z_0 \in \mathcal{Z}$ such that $x_0 = x_{\mu^0}^{z_0}$ and

$$\begin{aligned} J_0(\mu^0) &= \sup \{J(x, \mu^0), x \in X_{\mu^0}\} \\ &\leq \int_0^T \int_{\Gamma} l(t, x_0, \sigma) \mu_t^0(d\sigma) dt + \varepsilon. \end{aligned}$$

Then, by Theorem 3.1, for the fixed $z_0 \in \mathcal{Z}$, we can find a sequence $\{x_n\}$ such that $x_n = x_{\mu^n}^{z_0}$ and $x_n \rightarrow x_0$ in $C(\bar{I}, H)$. Recalling that every lower semicontinuous measurable integrand (see Condition L(2)) is the limit of an increasing sequence of Caratheodory integrands, we have

$$\begin{aligned} m \leq J_0(\mu^0) &\leq \int_0^T \int_{\Gamma} l(t, x_0, \sigma) \mu_t^0(d\sigma) dt + \varepsilon \\ &\leq \liminf \int_0^T \int_{\Gamma} l(t, x_n(t), \sigma) \mu_t^n(d\sigma) dt + \varepsilon \\ &\leq \liminf J_0(\mu^n) + \varepsilon \leq m + \varepsilon. \end{aligned}$$

Since ε is arbitrary, $J_0(\mu^0) = m$. This proves the theorem.

Remark. Suppose that $U: I \rightarrow 2^\Gamma / \{\emptyset\}$ is graph measurable, and replace $U(\mathcal{M})$ by $S_U(S_\Sigma)$ where $S_U = \{u \in U, u \in U(t) \text{ almost everywhere}\}$, $S_\Sigma = \{\mu \in \mathcal{M}, \mu_t(U(t)) = 1 \text{ almost everywhere}\}$ (see [7]). Then all the results of this paper remain valid.

Some comments on applications. (1) Consider a semilinear heat equation on an open bounded connected domain $\Omega \subset R^3$, with a Dirichlet boundary condition

$$\begin{aligned} \partial x(t, \xi) / \partial t &= \Delta x + F(x, u), \\ x(t, \xi) &= 0, \xi \in \partial\Omega, \\ x(0, \xi) &= x_0(\xi). \end{aligned}$$

Define the operator A by

$$D(A) \equiv \{\phi \in L_2(\Omega): \Delta\phi \in L_2(\Omega), \phi|_{\partial\Omega} = 0\};$$

and

$$A\phi \equiv \Delta\phi \text{ for } \phi \in D(A).$$

Take $E \equiv H_0^1$ and $E^* \equiv H^{-1}$ and $H \equiv L_2(\Omega)$. The operator A as defined above is the generator of a C_0 -semigroup of contractions in H ; but, under the Dirichlet boundary condition, it is a bounded operator from E to E^* and $-A$ is monotone (in fact strictly monotone). For $x \in E$, $u \in U$, define $f(x, u) \equiv F(x(\cdot), u)$. Then, under the given

assumptions, the results presented in the paper easily apply to this example. In fact, here we can take the sum $A + f$ as the operator f in our main theoretical results. This gives a regular but strongly nonlinear problem. Our emphasis here is on the additional singularity introduced by the presence of an unbounded nonpositive linear operator $A(t)$ appearing in the evolution equations (1) and (2). The results presented in the paper thus applies to more general situations as stated in the next example.

(2) Consider $-A$ to be a maximal monotone operator with domain in E and range in E^* and $-f$ a monotone operator satisfying Conditions F1–F5. Then the results presented in the paper apply to the problem

$$dx/dt = Ax + f(x, u)$$

with I satisfying Condition L. Note that here the operator A can absorb certain non-homogenous but fixed boundary conditions not included in the space E and hence may not be a bounded operator from E to E^* . In view of the remarks on the identity $L_s = L_w$, as discussed in § 3, our results obviously hold for any generator A of a C_0 -semigroup of contractions in H .

Acknowledgments. The authors thank the anonymous reviewers for their valuable comments.

REFERENCES

- [1] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, North-Holland, New York, 1981.
- [2] N. U. AHMED, *Properties of relaxed trajectories for a class of nonlinear evolution equations on a Banach space*, SIAM J. Control Optim., 21 (1983), pp. 953–967.
- [3] C. DELLACHERIE AND P. A. MEYER, *Probabilities and Potential*, North-Holland, Amsterdam, New York, 1978.
- [4] J. DIESTEL AND J. UHL, *Vector measures*, Math. Surveys Monoger, 15, American Mathematical Society, Providence, RI, 1977.
- [5] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [6] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math., Vol. 580, Springer-Verlag, Berlin, New York, 1977.
- [7] E. J. BALDER, *A general denseness result for relaxed control theory*, Bull Austral. Math. Soc., 30 (1984), pp. 463–475.
- [8] N. S. PAPAGEORGIOU, *Properties of the relaxed trajectories of evolution equations and optimal control*, SIAM J. Control Optim., 27 (1989), pp. 267–288.
- [9] N. U. AHMED, *An existence theorem for differential inclusions on Banach space*, J. Appl. Math. and Stochastic Anal., 5 (1992), pp. 123–130.
- [10] F. E. BROWDER, *Nonlinear initial value problems*, Ann. Math., 81 (1965), pp. 51–87.
- [11] ———, *Nonlinear equations of evolutions*, Ann. Math., 80 (1964), pp. 485–523.
- [12] ———, *Existence and uniqueness theorems for solutions of nonlinear boundary value problems*, Proc. Symp. Appl. Math., Vol. 17, American Mathematical Society, Providence, RI, 1965, pp. 24–29.
- [13] T. KATO, *Nonlinear Evolution Equations in Banach Spaces*, Proc. Symp. Appl. Math., Vol. 17, American Mathematical Society, Providence, RI, 1965, pp. 50–67.

ON THE GLOBAL DYNAMICS OF ADAPTIVE SYSTEMS: A STUDY OF AN ELEMENTARY EXAMPLE*

MARTÍN D. ESPAÑA† AND LAURENT PRALY‡

Abstract. The inherent nonlinear character of adaptive systems poses serious theoretical problems for the analysis of their dynamics. On the other hand, the importance of their dynamic behavior is directly related to the practical interest in predicting such undesirable phenomena as nonlinear oscillations, abrupt transients, intermittence or a high sensitivity with respect to initial conditions. A geometrical/qualitative description of the phase portrait of a discrete-time adaptive system with unmodeled disturbances is given. For this, the motions in the phase space are referred to normally hyperbolic (structurally stable) locally invariant sets. The study is complemented with a local stability analysis of the equilibrium point and periodic solutions. The critical character of adaptive systems under rather usual working conditions is discussed. Special emphasis is put on the causes leading to intermittence. A geometric interpretation of the effects of some commonly used palliatives to this problem is given. The “dead-zone” approach is studied in more detail. The predicted dynamics are compared with simulation results.

Key words. adaptive systems, intermittency, dynamic systems, discrete time systems, periodic solutions, invariant sets

AMS subject classifications. 93B27, 93C10, 93C40

1. Introduction. It is an already well-known fact that adaptive systems may exhibit very complicated dynamics. For instance, Anderson (1985), first showed that, although bounded, abrupt and explosive transients may occur in the presence of disturbances. Such undesirable behavior is not exclusive to adaptive control schemes inasmuch as it can occur in any system with parametric feedback (i.e., whose parameters are functions of the signals generated by the system itself). Other examples are output error and serial-parallel identification schemes. The nonlinear character of these systems poses serious theoretical problems for its dynamic analysis. However, from a practical point of view, a successful implementation is based on a thorough knowledge of the circumstances under which nonlinear oscillations, abrupt transients, or even intermittency may occur. The study of the sensitivity of the solutions with respect to the initial conditions is also of obvious importance given that, as pointed out by Bergé, Pomeau, and Vidal (1984), this circumstance happens to be intimately related with the existence of strange attractors (and chaos). The influence of external inputs on the overall behavior also needs careful attention. The above considerations have been at the origin of increasing interest during the past years on the dynamic description of adaptive systems and, more specifically, in the explanation of the occurrence of intermittent bursts. In the presence of stochastic disturbances, Anderson (1985), suggests that “bursting” is a consequence of nonpersistently exciting reference signals. However, as shown in this paper, intermittency may subsist and (as already shown by Narendra and Anaswamy (1986)) even solutions with unbounded parametric components may take place if the reference, although persistently exciting, has not enough energy. Jaïdane-Saïdane and Macchi (1988) have proposed a heuristic explanation to the intermittent phenomenon and attributed to it a “self-stabilizing” property implying bounded signals of the closed-loop adaptive linear system. The general validity of the last conclusion has, however, been already criticized by Egardt

* Received by the editors January 16, 1990; accepted for publication (in revised form) April 23, 1992.

† NASA-Ames/Dryden Flight Research Facility, P.O. Box 273, M/S D2002, Edwards, California 93523 (mespana@mcimail.com).

‡ Centre d'Automatique et Informatique, Ecole de Mines de Paris, 35 Rue Saint-Honoré, 77305 Fontainebleau Cedex, France (praly@caia.ensmp.fr).

(1979), who has established that bounded perturbations and reference signals may produce unbounded outputs unless the parameters remain bounded. More recently, intermittent bursts have been studied extrapolating a local analysis (around critical points such as equilibria and 2-periodic solutions) performed using bifurcation techniques (Golden and Ydstie (1988); Rey, Bitmead, and Johnson (1991)). Analysis based on averaging approximations are performed in Sethares and Mareels (1991) and España (1991). The latter shows that intermittency with either a continuous change (in average) or almost fixed parameter values may take place, the latter case being associated with self-oscillating modes. The use of averaging can rigorously be justified by the existence of the attractive locally invariant set. A concept is developed in detail in this paper for a particular example (see also Praly (1990)). We think, however, that the key issue to understand, and thus to prevent, undesired dynamics, is to address the problem of the global description of the trajectories in the phase space. To our knowledge, this is the first time that a global geometrical description of the phase portrait has been given for an adaptive system. Only local results have been obtained before. They are almost all contained in Ljung and Soderstrom (1983), Anderson et al. (1986), Riedle and Kokotovic (1986), and Benveniste, Metvier, and Priouret (1987). Although this analysis is particular to our example, more general conclusions can be obtained, since we use mathematical tools as integral sets (Praly (1985), (1990); Riedle and Kokotovic (1986)) or the existence of periodic solutions (Ljung (1977); Bodson et al. (1986); Pomet, Coron, and Praly (1990)), which have been shown to be applicable in more general situations. The analysis is made in a deterministic context and special emphasis is put on the causes leading to intermittent bursting.

2. Problem formulation. A sufficiently general formulation of a linear discrete-time system in closed loop with an adaptive controller is given by the following equations (see, for instance, Pomet, Coron, and Praly (1990)):

$$y(t+1) = A(\theta)y(t) + B(\theta)w(t), \quad \theta(t+1) = \theta(t) + \mu C(y(t), \theta(t), w(t)),$$

where μ is usually used to control the adaptation speed; $w(t)$ represents all the external inputs, including the output reference signal $r(t)$ and any unmodeled disturbances. The first equation is the regressor model and the second is the parameters updating algorithm.

In what follows, a discrete-time first-order plant with an adaptive proportional output-feedback controller is considered. The objective is to regulate the plant's output to a constant value r . Any possible mismatch between the model, used for the control purposes, and the plant is represented by the unknown and unmeasurable equation error $d(t)$:

$$(1) \quad d(t) = y(t) - ay(t-1) - u(t-1).$$

We refer to "the ideal case" when $d(t) \equiv 0$. Normally, $d(t)$ is the result of unmeasured disturbances or unmodeled dynamics. In our analysis, $d(t)$ is supposed constant. This particular situation arises in practice when, due to a (possible temporary) misalignment in the actuator, a bias exists in the effective control action applied.

A proportional controller has as an effect to shift the pole of this plant, and if the parameter a is known, any possible stable value can be assigned for it. When a is unknown, the following adaptive controller with a normalized gradient type updating parameter equation (Goodwin and Sin (1984), also called stochastic approximation by Egardt (1979)),

$$(2) \quad u(t) = -\theta(t)y(t) + r$$

$$(3) \quad \theta(t) = \theta(t-1) + \mu \frac{y(t-1)(y(t) - r)}{(1 + y(t-1)^2)}$$

guarantees, for the ideal case and for any $a \in \mathbb{R}$, that $y(t) \rightarrow r$ when $t \rightarrow \infty$. Moreover, if $r \neq 0$, $\theta(t) \rightarrow a$ as $t \rightarrow \infty$. The choice of a normalized-type algorithm (plus, perhaps, a uniform bound for the parameters not considered here) is crucial in practice to assure bounded signals, particularly when (bounded) disturbances are present (Egardt (1979)).

The system (1) in closed loop with the controller (2), (3), results in

$$\begin{aligned} y(t+1) &= -\psi(t)y(t) + d + r \\ (\Sigma_1) \quad \psi(t+1) &= \psi(t) + \mu \frac{y(t)(d - \psi(t)y(t))}{1 + y(t)^2}, \end{aligned}$$

where $\psi = \theta - a$. When d is nonzero, the change of variables $x = y/d$; $\psi = \psi$; $\alpha = r/d$, transforms (Σ_1) into

$$\begin{aligned} x(t+1) &= -\psi(t)x(t) + 1 + \alpha \\ (\Sigma) \quad \psi(t+1) &= \psi(t) + d^2\mu \frac{x(t)(1 - \psi(t)x(t))}{1 + d^2x^2(t)}. \end{aligned}$$

This variable rescaling puts in relief the role played by the reference-to-disturbance relationship α , called by Narendra and Annaswamy (1986) "persistent excitation of the reference relative to the disturbance." Moreover, it allows us to better describe the system's behavior for r and d close to zero. As discussed later, this (slightly disturbed regulation regime) is a very critical working condition. In the rescaled system, d^2 controls the adaptation speed of the algorithm. For our purposes we can thus assume that $\mu = 1$. The developments that follow can be done for the original system (Σ_1) replacing, when appropriate, the statement " d^2 sufficiently small" for " μ sufficiently small." In the first case, the slow adaptation condition is a consequence of the low level of the signals involved.

We can easily verify that (Σ) has a fixed point in $(\psi, x) = (1/\alpha, \alpha)$ if and only if α is nonzero and that it is unique if and only if $\alpha \neq -1$. In terms of the original system, this equilibrium corresponds to the output equal to the reference signal. The control objective is thus perfectly achieved at the fixed point.

The simulations show that for small values of d^2 , the behavior of the solutions of (Σ) is characterized by the following stages.

(a) Explosive stage: growth of the modulus of the x -component in the "instability" set $\{|\psi| > 1\}$.

(b) ReInjection stage: decrease of the modulus of the ψ -component in the set $\{|x| \text{ "large", } |\psi| > 1\}$ until $|\psi| < 1$.

(c) Implosive stage: decrease of the modulus of the x -component in the "stability" set $\{|\psi| < 1\}$.

(d) Drift-ejection stage: slow growth of $|\psi|$ leading a solution from the "stability" set to the "instability" set.

(e) When (d) does not occur, the desired working condition is globally attractive.

Stages (a)–(c) are very short in time and, under certain conditions, stage (d) may be performed very slowly. In such a case, two successive occurrences of stages (a)–(c) are separated by a very long period of time. The result is an intermittent phenomenon, as studied by Pomeau and Manville (1980), characterized by a succession of "bursts" on the x -component separated by long quiescent periods. In practice, some palliatives (such as dead zone or leakage (Egardt (1979)), normalization (Praly (1983)), internal model principle (Elliott and Goodwin (1984)), filtering (Anderson et al. (1986)), etc.) are used to avoid intermittency and other nondesirable behaviors. However, if these remedies are not appropriately chosen, a qualitatively similar behavior may be observed for these more intricate cases (see Praly (1988)). The effect of some of these modifications

in our example is discussed in §§ 6 and 7. To give a geometrical explanation of stages (a)–(e), the existence of two locally invariant (under the action of (Σ)) sets is demonstrated using the graph transform technique (Shub (1987)). The first one is repellent, explaining stage (a). The second one is attractive and allows us to explain stages (c), (d), and (e). Finally, stage (b) results from (a), when, during bursts, the disturbance d becomes negligible with respect to the x -component of the solutions. These locally invariant graphs are easily computed when the ψ -component remains constant. For this, we consider the set of all bounded solutions of (Σ) when $d = 0$ given by

$$(4) \quad S_0 \equiv \{(\psi, h(\psi)) \in \mathbb{R}^2 / h(\psi) = (1 + \alpha)/(1 + \psi), \psi \neq -1\}.$$

S_0 , called the “frozen parameters invariant set,” is invariant under the map $\Sigma_{d=0}$ (i.e., $\Sigma_{d=0}(S_0) \subseteq S_0$) and has exactly the properties associated with (a) and (c). It seems reasonable to expect that, when $|d|$ is not zero but still small, locally invariant graphs, approximated by S_0 , still exist. The idea of using locally invariant sets or, more generally, locally integral sets, has been introduced by Riedle and Kokotovic (1986) and Praly (1985), (1990). However, their existence was only established for the “stability” set $\{|\psi| < 1\}$ and locally with respect to the x -components.

The paper is organized as follows. In § 3, the existence and properties of locally invariant sets are established. Critical elements and locally invariant sets are combined in § 4 to obtain theoretical results on the system global dynamics. These results are interpreted and compared with simulations in § 5. In § 6, the effects of introducing a “dead-zone” in the algorithm (3) is discussed. Finally, § 7 is dedicated to our concluding remarks. The critical elements—fixed points and periodic solutions—of (Σ) and the corresponding nearby local behavior needed to complement the global analysis are considered in the Appendix.

3. Locally invariant sets. In § 2, we observed that for $d = 0$, the set S_0 is invariant under $\Sigma_{d=0}$. Now from the definition of (Σ) , when $\psi(t)$ and $x(t)$ are such that $1 + \psi(t)$ and $1 + \psi(t) + d^2 x(t)(1 + x(t))$ are nonzero we have

$$\begin{aligned} \left(x(t+1) - \frac{1 + \alpha}{1 + \psi(t+1)} \right) &= -\psi(t) \left(x(t) - \frac{1 + \alpha}{1 + \psi(t)} \right) \\ &\quad + \frac{d^2(1 + \alpha)x(t)(x(t)\psi(t) - 1)}{(1 + \psi(t))(1 + \psi(t) + d^2 x(t)(1 + x(t)))}. \end{aligned}$$

The presence of d^2 in the second term on the right-hand side shows that S_0 is close to being a locally invariant set of (Σ) with d nonzero. Finally, for $d = 0$, this expression proves that (i) $S_0 \cap \{(\psi, x) | |\psi| > 1\}$ is exponentially repellent and (ii) $S_0 \cap \{(\psi, x) | |\psi| < 1\}$ is exponentially attractive. These remarks lead us to look for locally invariant sets close to S_0 , which are repellent in the set $\{|\psi| > 1\}$ and attractive in the set $\{|\psi| < 1\}$. These sets will be used as references for the global description of the solutions of (Σ) .

3.1. The repellent locally invariant set (RLIS). Given any nonzero d , let ε be the smallest positive root of

$$(5) \quad \Delta(\varepsilon) = \left(\varepsilon - \frac{|d|}{1 + \varepsilon} \right) - 2 \sqrt{\frac{|1 + \alpha||d|(1 + \varepsilon + |d|)}{\varepsilon}}.$$

For any function $M: \{|\psi| \geq 1 + \varepsilon\} \rightarrow \mathbb{R}$, we define its image by the operator T as

$$(6) \quad TM(\psi) = \frac{1 + \alpha - M(\phi_M(\psi))}{\psi},$$

where ϕ_M is defined in terms of the function

$$(7) \quad \hat{\phi}_M(\psi) = \psi + d^2 M(\psi) \frac{(1 - \psi M(\psi))}{1 + d^2 M^2(\psi)},$$

as follows:

$$(8) \quad \phi_M(\psi) = \begin{cases} \sup(1 + \varepsilon, \hat{\phi}_M(\psi)) & \text{if } \psi \geq 1 + \varepsilon, \\ \inf(-1 - \varepsilon, \hat{\phi}_M(\psi)) & \text{if } \psi < -1 - \varepsilon. \end{cases}$$

By definition, $\phi_M: \{|\psi| \geq 1 + \varepsilon\} \rightarrow \{|\psi| \geq 1 + \varepsilon\}$ is a continuous function and $\psi\phi_M(\psi)$ is positive. We are interested in the operator T because, if it has a fixed point H , then H satisfies the local invariance property:

$$(9) \quad H(\hat{\phi}_H(\psi)) = 1 + \alpha - \psi H(\psi) \quad \text{if } \psi \hat{\phi}_H(\psi) > 0, \quad |\psi| > 1 + \varepsilon, \quad |\hat{\phi}_H(\psi)| > 1 + \varepsilon.$$

The graph $\{(\psi, H(\psi)) / |\psi| \geq 1 + \varepsilon\}$ has two connected components in the plane (ψ, x) . They are such that, with its initial condition in one of these sets, any solution of (Σ) will stay in it unless its ψ -component leaves its corresponding definition interval $\{\psi > 1 + \varepsilon\}$ or $\{\psi < -(1 + \varepsilon)\}$. To exhibit the fixed point H , we consider the subset of $C^0(\{|\psi| \geq 1 + \varepsilon\}, \mathbb{R})$:

$$\mathcal{B} = \{M | \partial(M, 0) \leq m_0$$

$$\text{and } \text{sgn}(\psi_1) = \text{sgn}(\psi_2) \Rightarrow |M(\psi_1) - M(\psi_2)| \leq m_1 |\psi_1 - \psi_2|\}.$$

It is a complete metric space with the distance ∂ defined as

$$\partial(M_1, M_2) = \sup_{\{|\psi| \geq 1 + \varepsilon\}} |M_1(\psi) - M_2(\psi)|.$$

The constants m_0, m_1 are

$$(10) \quad m_0 = \frac{|1 + \alpha|}{\varepsilon}, \quad m_1 = \frac{2|1 + \alpha|}{\varepsilon(\varepsilon - |d|/1 + \varepsilon)}.$$

The next lemma and the uniform contraction theorem (see Hale (1980)) allow us to prove that T has a fixed point in \mathcal{B} .

LEMMA 1. For any nonzero d , let ε be given by (5), then (i) T maps \mathcal{B} into \mathcal{B} ; (ii) For $M_i, i = 1, 2$, in \mathcal{B} , we have $\partial(TM_1, TM_2) \leq \tau \partial(M_1, M_2)$, with

$$\tau = \frac{1}{1 + \varepsilon} + |d|m_1 \left(1 + \frac{|d|}{1 + \varepsilon}\right) < 1.$$

Proof. From (6) and definition (10), it can be easily shown that for all $M \in \mathcal{B}$, $\partial(TM, 0) \leq m_0$. The rest of the proof is obtained by showing that for all $M_1, M_2 \in \mathcal{B}$, if $\text{sgn}(\psi_1) = \text{sgn}(\psi_2)$ then $|TM_1(\psi_1) - TM_2(\psi_2)| \leq m_1 |\psi_1 - \psi_2| + \tau \partial(M_1, M_2)$. For this, use is made of (8) to write $|\hat{\phi}_M(\psi_1) - \hat{\phi}_M(\psi_2)| \geq |\phi_M(\psi_1) - \phi_M(\psi_2)|$. The details can be consulted in España and Praly (1988). \square

Another important property of H is the following. Let $(\psi_0, x_0) \in \{|\psi_0| > 1 + \varepsilon\} \times \mathbb{R}$. We denote by $(\psi_1, x_1) = \Sigma(\psi_0, x_0)$ its image by (Σ) . Suppose that $\text{sgn}(\psi_1) =$

$\operatorname{sgn}(\psi_0)$ and $|\psi| > 1 + \varepsilon$. With Lemma 1, the definition (7), (8), and the property (9) we have

$$\begin{aligned}
 |x_0 - H(\psi_0)| &\leq \left| \frac{H(\psi_1) - x_1}{\psi_0} \right| + \left| \frac{H(\phi_H(\psi_0)) - H(\psi_1)}{\psi_0} \right| \\
 &\leq \left| \frac{H(\psi_1) - x_1}{\psi_0} \right| + m_1 \left| \frac{\hat{\phi}_H(\psi_0) - \psi_1}{\psi_0} \right| \\
 (11) \quad &\leq \left| \frac{H(\psi_1) - x_1}{\psi_0} \right| + \frac{m_1 d^2}{|\psi_0|} \sup_x \left\{ \left| \frac{\partial}{\partial x} \left(\frac{x(1 - \psi_0 x)}{1 + d^2 x^2} \right) \right| \right\} |H(\psi_0) - x_0| \\
 &\leq \left| \frac{H(\psi_1) - x_1}{\psi_0} \right| + m_1 |d| \left(1 + \frac{|d|}{1 + \varepsilon} \right) |H(\psi_0) - x_0| \\
 &\leq \frac{1}{1 + (1 + \varepsilon)(1 - \tau)} |H(\psi_1) - x_1|.
 \end{aligned}$$

Hence, since τ is strictly smaller than 1, the distance from a solution to its projection, parallel to the x -axis, on the graph of H , must increase as long as its ψ -component stays in the same interval $\{\psi > 1 + \varepsilon\}$ or $\{\psi < -(1 + \varepsilon)\}$. Using the above derivations and the definitions (5), (10) it can be shown that

$$(12) \quad |H(\phi_H(\psi_0)) - H(\psi_1)| \leq \frac{m_1 |d| (1 + |d|/(1 + \varepsilon))}{1 - m_1 |d| (1 + |d|/(1 + \varepsilon))} |x_1 - H(\psi_1)|.$$

If we replace in (12) the value of m_1 given by (10) and use (5) again, we obtain

$$0 < m_1 |d| (1 + |d|/(1 + \varepsilon)) < \frac{1}{2} \Rightarrow |H(\phi_H(\psi_0)) - H(\psi_1)| < |x_1 - H(\psi_1)|.$$

Then from the following product, computed using the definition of (Σ) and the invariance property of H ,

$$(x_0 - H(\psi_0))(x_1 - H(\psi_1)) = -\frac{(H(\psi_1) - x_1)^2}{\psi_0} + \frac{(H(\phi_H(\psi_0)) - H(\psi_1))(x_1 - H(\psi_1))}{\psi_0}$$

we have $\operatorname{sgn}[(x_0 - H(\psi_0))(x_1 - H(\psi_1))] \neq \operatorname{sgn}(\psi_0)$. To summarize, we have established the following theorem.

THEOREM 1 (the RLIS). *For any nonzero d , let ε be given by (5). There exist a bounded Lipschitz continuous function H , defined on $\{|\psi| \geq 1 + \varepsilon\}$, with bound m_0 and a Lipschitz constant m_1 given by (10), such that*

(i) *If $|\psi| \geq 1 + \varepsilon$, $|\hat{\phi}_H(\psi)| \geq 1 + \varepsilon$, $\operatorname{sgn}(\hat{\phi}_H(\psi)) = \operatorname{sgn}(\psi)$, then*

$$(13) \quad H(\hat{\phi}_H(\psi)) = 1 + \alpha - \psi H(\psi).$$

(ii) *There exists ρ positive such that $(\psi, x) \in \{|\psi| \geq 1 + \varepsilon\} \times \mathbb{R}$, $(\phi, y) := \Sigma(\psi, x) \in \{|\psi| \geq 1 + \varepsilon\} \times \mathbb{R}$, and $\psi\phi > 0$ imply*

$$(14) \quad \operatorname{sgn}((x - H(\psi))(y - H(\phi))) \neq \operatorname{sgn}(\psi),$$

$$(15) \quad |y - H(\phi)| \geq (1 + \rho)|x - H(\psi)|.$$

(iii) *Approximation of $H : \sup_{\{|\psi| \geq 1 + \varepsilon\}} |H(\psi) - h(\psi)|/d^2$ is bounded when $d^2 \rightarrow 0$.*

Proof. Statements (i) and (ii) are already established. To prove (iii), we first note that h , defined in (4), belongs to \mathcal{B} . Then, using Lemma 1, we have $\partial(h, H) \leq \partial(h, Th) + \partial(Th, TH) \leq \partial(h, Th)/(1 - \tau)$. The result is finally obtained using the

definition of h and T to show that $|h(\psi) - Th(\psi)| = O(d^2)$. (The details can be consulted in España and Praly (1988).) \square

Remarks. 1. With (i) and (ii), this theorem establishes the existence of a globally, exponentially RLIS, given by the graph of a bounded continuous function $H : \{|\psi| > 1 + \varepsilon\} \rightarrow \mathbb{R}$ which, following (iii), can be approximated by the “frozen parameter invariant set,” for $|d|$ sufficiently small. It is important to emphasize at this point that the repulsiveness of this graph, expressed by (15), has a global character in the sense that it is valid for any starting point $x \in R$. This is a direct consequence of the use of a normalized updating parameter equation (see (3) and (11)).

2. According to the sign of its ψ -component, the x -component of any solution changes side or not with respect to the graph of H (see (14)).

3. Although T has a unique fixed point H in \mathcal{B} , H need not be the unique function in \mathcal{B} satisfying (13). This nonuniqueness comes from the arbitrariness of the function ϕ_M , which is not determined by (Σ) (see the discussion about the stopping function in Praly (1990)). In § 4 (see Theorem 3(b) and related remarks) the conditions under which the whole RLIS or a portion of it is unique are established.

3.2. The attractive locally invariant set (ALIS). Let

$$(16) \quad d^{*2} = \frac{1}{2|1 + \alpha|^2} \left(\sqrt{\frac{1 + 3|1 + \alpha|}{1 + 2|1 + \alpha|}} - 1 \right).$$

Taking $|d|$ in $(0, d^*)$, let η be the smallest positive root of

$$(17) \quad \Delta(\eta) = \left(\eta - \frac{d^2 n_0^2}{1 + d^2 n_0^2} \right) - 2 \sqrt{\frac{(1 + 2n_0)n_0 d^2}{1 + d^2 n_0^2}},$$

where n_0 is defined by

$$(18) \quad n_0 = \frac{|1 + \alpha|}{\eta}.$$

The constraint introduced on d assures that $\Delta(0)\Delta(1)$ is strictly negative. This implies that η is strictly smaller than 1. Now for any d , $0 < |d| < d^*$, we define an operator P acting on functions $N : \mathbb{R} \rightarrow \mathbb{R}$ by

$$(19) \quad PN(\phi) = \begin{cases} 1 + \alpha - \psi_N(\phi)N(\psi_N(\phi)) & \text{if } |\phi| \leq 1 - \eta, \\ PN(1 - \eta) \text{ (resp., } PN(\eta - 1)) & \text{if } \phi \geq 1 - \eta \text{ (resp., } \leq -(1 - \eta)), \end{cases}$$

where $\psi_N(\phi)$, mapping $\{|\psi| \leq 1 - \eta\}$ into $\{|\psi| \leq 1 - \eta\}$, is a function implicitly defined by the next two relations:

$$(20) \quad \phi = \frac{\hat{\psi}_N(\phi) + d^2 N(\hat{\psi}_N(\phi))}{1 + d^2 N^2(\hat{\psi}_N(\phi))},$$

$$(21) \quad \psi_N(\phi) = \begin{cases} \hat{\psi}_N(\phi) & \text{if } |\hat{\psi}_N(\phi)| \leq 1 - \eta, \\ (1 - \eta) \text{ (resp., } -(1 - \eta)) & \text{if } \hat{\psi}_N(\phi) > 1 - \eta \text{ (resp., } < -(1 - \eta)). \end{cases}$$

We are interested in the operator P because, if it has a fixed point G , then G satisfies the local invariance property:

$$(22) \quad G(\phi) = -\hat{\psi}_G(\phi)G(\hat{\psi}_G(\phi)) + 1 + \alpha, \text{ if } |\phi| \leq 1 - \eta \text{ and } |\hat{\psi}_G(\psi)| \leq 1 - \eta.$$

As for H in the “instability” set, the graph $\{(\psi, G(\psi))/|\psi| \leq 1 - \eta\}$ defines a set in the plane (ψ, x) . It is such that, with its initial condition in this set, any solution of (Σ) will stay in it unless its ψ -component leaves the set of “strict stability” $\{|\psi| \leq 1 - \eta\}$.

To exhibit the fixed point G , we consider the set of “saturated” functions:

$$(23) \quad \mathcal{C} = \left\{ N \in C^0(\mathbb{R}, \mathbb{R}) \left| \begin{array}{l} (1) \partial(N, 0) \leq n_0 \\ (2) \forall \psi_1, \psi_2 \in \mathbb{R}, |N(\psi_1) - N(\psi_2)| \leq n_1 |\psi_1 - \psi_2| \\ (3) N(\psi) = \begin{cases} N(1 - \eta) & \text{if } \psi \geq 1 - \eta \\ N(-(1 - \eta)) & \text{if } \psi \leq -(1 - \eta) \end{cases} \end{array} \right. \right\},$$

$$(24) \quad n_1 := \sqrt{(1 + d^2 n_0^2) n_0 / (1 + 2n_0) d^2},$$

which is a complete metric space with the distance

$$\partial(N_1, N_2) := \sup_{\mathbb{R}} \{|N_1(\psi) - N_2(\psi)|\}.$$

Before studying the operator P acting on \mathcal{C} , we must be sure that $\hat{\psi}_N(\phi)$, implicitly defined by (20), makes sense. For this we have the following lemma.

LEMMA 2. *For any d , $0 < |d| < d^*$ and η given by (17), there exists a function $D: \mathcal{C} \times \{|\phi| \leq 1 - \eta\} \rightarrow \mathbb{R}$ and positive numbers $b_\phi(n_0, n_1, d^2)$, $b_n(n_0, n_1, d^2)$ satisfying for (N_i, ϕ_i) in $\mathcal{C} \times \{|\phi| \leq 1 - \eta\}$ and $i = 1, 2$*

$$(25) \quad |D(N_1, \phi_1) - D(N_2, \phi_2)| \leq b_\phi |\phi_1 - \phi_2| + b_n \partial(N_1, N_2),$$

$$(26) \quad |D(N_i, \phi_i)| \leq n_0(1 + n_0),$$

$$(27) \quad D(N, \phi) = N(\phi - d^2 D)(1 - \phi N(\phi - d^2 D)).$$

Proof. The proof takes advantage of the “almost identity” character of the function defined by (20) and (21) when d^2 is sufficiently small. Moreover, it gives conditions on the sizes of d^2 and η (see (16), (17)). The details may be consulted in España and Praly (1988); a more general result is also established in Praly (1990). \square

With this function D , we can rewrite (20) as follows:

$$(28) \quad \hat{\psi}_N(\phi) = \phi - d^2 D(N, \phi).$$

The next lemma and the uniform contraction theorem (see Hale (1980)) allow us to prove that P has a fixed point in \mathcal{C} .

LEMMA 3. *For any d , $0 < |d| < d^*$ and η given by (17), we have (i) P maps \mathcal{C} into \mathcal{C} . (ii) For all N_i , $i = 1, 2$, in \mathcal{C} , we have $\partial(PN_1, PN_2) \leq \lambda \partial(N_1, N_2)$ with $\lambda < 1$.*

Proof. It can be easily checked using (18)–(21) that PN satisfies (1) and (3) of (23). By using the properties of N , D , and the fact that $|\phi| \leq 1 - \eta$, we obtain

$$(29) \quad \begin{aligned} |PN_1(\phi_1) - PN_2(\phi_2)| &\leq \lambda(\eta, d^2) \partial(N_1, N_2) + n_1^* |\phi_1 - \phi_2|, \\ n_1^* &:= ((1 - \eta)n_1 + n_0)(1 + d^2 b_\phi) \leq n_1. \end{aligned}$$

Now, thanks to the choice of η in (17) we can show that

$$(30) \quad \lambda(\eta, d^2) := (1 - \eta)(1 + d^2 b_n n_1) + n_0 b_n d^2 = \frac{1 - (1 + 2n_0)d^2 n_1}{1 + d^2 n_0^2} < 1.$$

The details can be found in España and Praly (1988). See also Lemma 1 of Praly (1990). \square

We next establish an important feature of the graph of G with respect to the solutions of (Σ) . Let (ψ_0, x_0) be an element of $\{|\psi| \leq (1 - \eta)\} \times \{|x| \leq \xi\}$ and (ψ_1, x_1) its image by (Σ) . Whenever $|\psi_1|$ is smaller than $1 - \eta$, from (28), (22), (19) we have

$$\begin{aligned} |x_1 - G(\psi_1)| &\leq |\psi_0| |G(\psi_0) - x_0| + |\psi_0| |G(\psi_G(\psi_1)) - G(\psi_0)| \\ &\quad + |G(\psi_G(\psi_1))| |\psi_G(\psi_1) - \psi_0| \\ &\leq |\psi_0| |G(\psi_0) - x_0| + (n_0 + |\psi_0| n_1) |\hat{\psi}_G(\psi_1) - \psi_0|. \end{aligned}$$

On the other hand, adding and subtracting $G(\psi_G)(1 - \psi_1 G(\psi_0))$ we obtain

$$\begin{aligned} |\hat{\psi}_G(\psi_1) - \psi_0| &= d^2 |x_0(1 - x_0 \psi_1) - G(\hat{\psi}_G(\psi_1))(1 - \psi_1 G(\hat{\psi}_G(\psi_1)))| \\ &\leq d^2(1 + \xi + n_0) |x_0 - G(\psi_0)| + d^2(1 + 2n_0)n_1 |\psi_0 - \hat{\psi}_G(\psi_1)| \\ &\leq \frac{d^2(1 + \xi + n_0)}{1 - d^2(1 + 2n_0)n_1} |x_0 - G(\psi_0)|. \end{aligned}$$

Hence, using (17) and (24), we have established

$$(31) \quad |x_1 - G(\psi_1)| \leq \sigma(\xi) |x_0 - G(\psi_0)|,$$

$$(32) \quad \sigma(\xi) = \frac{1 - (1 + 2n_0)d^2 n_1}{1 + d^2 n_0^2} + \frac{d^2 n_1}{1 + d^2 n_0^2} (\xi - n_0).$$

With (30), $\sigma(\xi) > 0$ is strictly smaller than 1 if

$$(33) \quad n_0 < \xi < n_0 + \frac{n_0^2 + n_1(1 + 2n_0)}{n_1}.$$

Thus, any solution staying in the set $\{|\psi| \leq (1 - \eta)\} \times \{|x| \leq \xi\}$, with ξ satisfying (33), exponentially approaches the graph of G . Moreover, with the above derivations, we have

$$|\psi_G(\psi_1)G(\psi_G(\psi_1)) - \psi_0 G(\psi_0)| \leq \frac{d^2(1 + \xi + n_0)n_1}{1 + d^2 n_0^2} |x_0 - G(\psi_0)|.$$

However, since the invariance property of G implies

$$\begin{aligned} \psi_0(x_1 - G(\psi_1))(x_0 - G(\psi_0)) &= -\psi_0^2(x_0 - G(\psi_0))^2 + \psi_0(x_0 - G(\psi_0)) \\ &\quad \cdot (\psi_G(\psi_1)G(\psi_G(\psi_1)) - \psi_0 G(\psi_0)), \end{aligned}$$

it follows that from $|x_0| \leq \xi$ and $d^2(1 + \xi + n_0)n_1/(1 + d^2 n_0^2) \leq |\psi_0| \leq 1 - \eta$ we obtain $\text{sgn}[(x_1 - G(\psi_1))(x_0 - G(\psi_0))] \neq \text{sgn}(\psi_0)$. To summarize, we have established the following theorem.

THEOREM 2 (the ALIS). *For any d , $0 < |d| < d^*$, and η given by (17), there exists a bounded Lipschitz continuous function G with bound n_0 and Lipschitz constant n_1 given, respectively, by (18) and (24), such that*

(i) *If $|\phi| \leq 1 - \eta$ and $|\hat{\psi}_G(\phi)| \leq 1 - \eta$, then*

$$(34) \quad G(\phi) = 1 + \alpha - \psi G(\psi) \quad \text{and} \quad \phi = \psi + d^2 \frac{G(\psi)(1 - \psi G(\psi))}{1 + d^2 G(\psi)^2}.$$

(ii) *Let ξ satisfy*

$$n_0 < \xi < n_0 + \frac{n_0^2 + n_1(1 + 2n_0)}{n_1},$$

then there exists $\sigma(\xi) < 1$ such that $(\psi, x) \in \{|\psi| < 1 - \eta\} \times \{|x| \leq \xi\}$ and $(\phi, y) := \Sigma(\psi, x) \in \{|\phi| \leq 1 - \eta\} \times \mathbb{R}$ imply

$$(35) \quad |y - G(\phi)| \leq \sigma(\xi)|x - G(\psi)|.$$

Moreover, if

$$\frac{d^2(1 + \xi + n_0)n_1}{1 + d^2n_0^2} \leq |\psi|,$$

then $\text{sgn}((y - G(\phi))(x - G(\psi))) \neq \text{sgn}(\psi)$.

(iii) Approximation of G : $\sup_{\{|\psi| < (1 - \eta)\}} |G(\psi) - h(\psi)|/d^2$ is bounded for $d^2 \rightarrow 0$.

Proof. Statement (i) is a direct consequence of Lemma 3. Statement (ii) follows from (31). To prove (iii), we first note that h , defined in (4), belongs to \mathcal{C} . Now, since G is the fixed point of P , Lemma 3(ii) gives $\partial(h, G) \leq \partial(h, Ph)/(1 - \lambda(\eta, d^2))$. The result is finally obtained using the definition of h and P to show that (see details in España and Praly (1988))

$$(36) \quad \partial(h, G) \leq d^2n_0(n_0 + n_1)(n_0 + 1)/(1 - \lambda(\eta, d^2)). \quad \square$$

Remarks. 1. This theorem establishes the existence of an (exponentially) ALIS given by the graph of a bounded continuous function $G: \{|\psi| < 1 - \eta\} \rightarrow \mathbb{R}$ which, following (iii), can be approximated by the “frozen-parameter invariant set,” when $|d|$ is sufficiently small.

2. If its ψ -component is larger than

$$\frac{d^2(1 + \xi + n_0)n_1}{1 + d^2n_0^2},$$

respectively, smaller than

$$-\frac{d^2(1 + \xi + n_0)n_1}{1 + d^2n_0^2},$$

the x -component of any solution changes side with respect to (respectively, remains on the same side of) the graph of G .

3. Even though G is the unique fixed point of P in \mathcal{C} , its graph need not be the only one satisfying (22). The nonuniqueness of the ALIS comes from the arbitrariness of the definition (21), which is not determined by (Σ) .

3.3. Additional characterization of the locally invariant sets. The existence Theorems 1 and 2 do not give enough information about the location of the locally invariant sets in the phase space. For this we have the following useful property.

Property 1. In their respective domains of definition, the functions H and G , determined by Theorems 1 and 2, belong to the family of functions F_α whose elements satisfy

(i) If $1/\alpha < -1$, then

$$\psi < 1/\alpha \Rightarrow M(\psi)(1 - \psi M(\psi)) \leq 0, \quad \text{and} \quad \psi > 1/\alpha \Rightarrow M(\psi)(1 - \psi M(\psi)) \geq 0.$$

(ii) If $1/\alpha > -1$, then

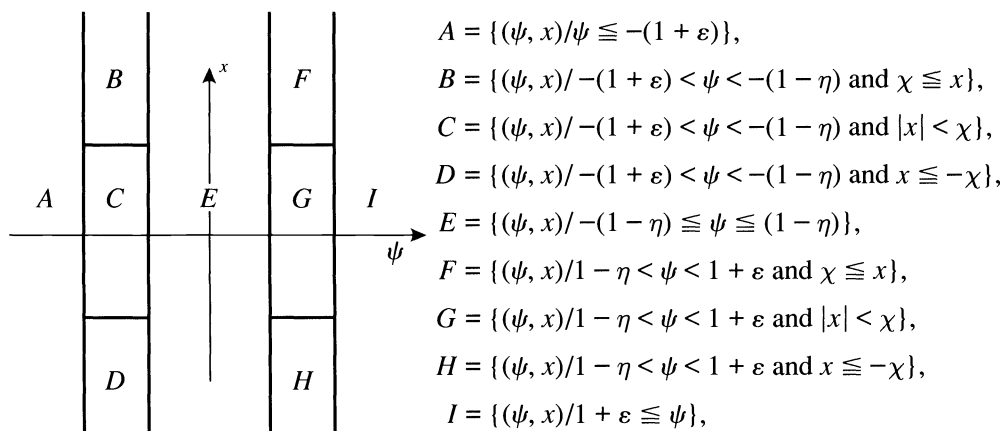
$$\psi < 1/\alpha \Rightarrow M(\psi)(1 - \psi M(\psi)) \geq 0, \quad \text{and} \quad \psi > 1/\alpha \Rightarrow M(\psi)(1 - \psi M(\psi)) \leq 0.$$

(iii) If $\alpha = -1 \Rightarrow M(\psi) \equiv 0$.

Proof. Statements (i) and (ii) are demonstrated by showing that, in the corresponding domain of definition, F_α is a closed subset of \mathcal{B} (respectively, \mathcal{C}) such that $TF_\alpha \subseteq$

F_α (respectively, $PF_\alpha \subseteq F_\alpha$) (see España and Praly (1988)). Note that F_α does not depend on d^2 . Statement (iii) results from a continuity argument.

4. Global behavior of the solutions: Technical results. Knowing the existence of critical elements and locally invariant sets, we are now in position for studying the behavior of the solutions. We decompose the plane (ψ, x) into nine subsets:



with ε given by (5), η by (17), and $\chi > 1/(1 - \eta)$. The global behavior of the solutions can be understood by looking at their evolution in each of these sets on the locally invariant graphs and outside them. We call $A \cup I$ the “strict instability” set, E the “strict stability” set, and $B \cup C \cup D \cup F \cup G \cup H$ the “critical stability” set.

4.1. Solutions in the locally invariant sets (RLIS and ALIS).

THEOREM 3(a) (the stationary solution). (i) For any nonzero d and ε given by (5), the (unique) equilibrium point of (Σ) belongs to the RLIS if and only if $|1/\alpha| \geq 1 + \varepsilon$.

(ii) For any d , $0 < |d| < d^*$, and with d^* , η given by (16), (17), the (unique) equilibrium point of (Σ) is in the ALIS if and only if $|1/\alpha| \leq 1 - \eta$.

(iii) When $\alpha = -1$, $H(\psi) \equiv 0$, $G(\psi) \equiv 0$ and any point in the RLIS or ALIS is an equilibrium point.

Proof. Given the global repulsiveness (respectively, attractiveness) of the RLIS (respectively, ALIS), the fixed point must be in the RLIS (respectively, ALIS) if it is in $\{|\psi| > 1 + \varepsilon\}$ (respectively, $\{|\psi| < 1 - \eta\}$). The rest follows from Theorem A in the Appendix and Property 1. \square

THEOREM 3(b) (the nonstationary solutions). If $\psi(t)$ is the ψ -component of any solution of (Σ) with initial condition in a locally invariant set, then

(i) if $1/\alpha > -1 \Rightarrow (\psi(t) - 1/\alpha)/(\psi(t+1) - 1/\alpha) > 1$.

(ii) if $1/\alpha < -1 \Rightarrow (\psi(t+1) - 1/\alpha)/(\psi(t) - 1/\alpha) > 1$.

Proof. Theorem 3(b) is a consequence of Property 1 and the definition of (Σ) . \square

Remarks. From Theorems 3(a) and 3(b), if $1/\alpha < 109-1$ (respectively, $1/\alpha > -1$), the ψ -component of the solutions on any locally invariant set moves monotonically away from (respectively, toward) the value $1/\alpha$.

1. If $|1/\alpha| > 1$ and $(\psi, x) \in \text{RLIS}$ with $\psi \in (-\infty, 1/\alpha) \cup (1 + \varepsilon, \infty)$, then it can be seen, with (8), that $\phi_H(\psi) = \hat{\phi}_H(\psi)$. Hence, the portion of the RLIS defined in $(-\infty, 1/\alpha) \cup (1 + \varepsilon, \infty)$ is unique since the “stopping mechanism” is not active here. Note that the whole RLIS is unique if $1/\alpha > 1 + \varepsilon$; moreover, in this case, the RLIS and the stable manifold of the fixed point coincide over $\{|\psi| > 1 + \varepsilon\}$ (see Iooss (1979)).

2. If $1/\alpha < -1$ the ψ -component of the trajectories in the RLIS with $\psi(0) \in (-\infty, 1/\alpha) \cup (1/\alpha, \infty)$ are asymptotically unbounded.

3. If the fixed point lies in the ALIS (i.e., $|1/\alpha| < 1 - \eta$), it is a global attractor inside the ALIS. If it lies in the RLIS, the solutions in the ALIS leave the set $\{|\psi| < 1 - \eta\}$ in a finite time through the boundary $\psi = 1 - \eta$. An estimation of the “traveling” speed of the ψ -component in the ALIS for this case will be of interest in our analysis and is given by the next theorem.

THEOREM 3(c). *If the fixed point is not in the ALIS, the solutions in it leave the “strict stability” set through the boundary $\psi = 1 - \eta$ in a finite time.*

Proof. From Property 1, $\psi G(\psi) < 1$ and $G(\psi) > 0$. Now, G being continuous on the compact set $\{|\psi| \leq 1 - \eta\}$, there exists η_1 , strictly positive, such that: $G(\psi) > \eta_1$ and $1 - \psi G(\psi) > \eta_1$. With the definition of (Σ) this implies that for all ψ ; $|\psi| < 1 - \eta$,

$$(37) \quad n_0(1 + n_0) > \frac{\phi - \psi}{d^2} = \frac{G(\psi)(1 - \psi G(\psi))}{1 + d^2 G(\psi)^2} > \frac{\eta_1^2}{1 + d^2 n_0^2}.$$

Therefore, in the ALIS, $\psi(t)$ moves with positive speed of the order of d^2 , thus leaving the set $\{|\psi| < 1 - \eta\}$ in a finite time through the boundary $\psi = 1 - \eta$. \square

4.2. Solutions in the “strict instability” set outside the RLIS.

THEOREM 4. *For any $d \neq 0$ and with ε given by (5), we have*

(i) *Global repulsiveness.* While the solution remains in $\{\psi \geq 1 + \varepsilon\}$ (respectively, $\{\psi \leq -(1 + \varepsilon)\}$), it exponentially diverges from the RLIS crossing it at each time t (respectively, remaining on the same side of the RLIS).

(ii) *Injection.* If for some time t_0 , a solution satisfies $|\psi(t_0)| \geq 1 + \varepsilon$, $x(t_0) \neq H(\psi(t_0))$, then there exists a finite time $t_1 > t_0$, such that $|\psi(t_1)| < 1 + \varepsilon$. Hence, there is no solution satisfying every $x \neq H(\psi)$ and $|\psi| \geq 1 + \varepsilon$.

Proof. Statement (i) is a direct consequence of (14), (15). To prove (ii), we first note that if $|x| > 1$ and $|\psi| > 1 + \varepsilon$ then

$$(38) \quad 0 < \frac{d^2 x^2 + |\psi x|}{(1 + d^2 x^2)|\psi x|} < 1 - \frac{\varepsilon d^2}{(1 + \varepsilon)(1 + d^2)} < 1.$$

Hence, by Theorem 1, if for all s in $[t_0, t]$, $\psi(s) \geq 1 + \varepsilon$ (respectively, $\leq -(1 + \varepsilon)$) then

$$(39) \quad |x(t) - H(\psi(t))| \geq (1 + \rho)^{t-t_0} |x(t_0) - H(\psi(t_0))|.$$

Now since $H(\psi)$ is bounded, there exists a first time t_1 (depending on $x(t_0)$, $\psi(t_0)$) such that either $|\psi(t_1)| \leq 1 + \varepsilon$ or $|x(t_1)| \geq 1$. In the latter case, from (38) and the definition of (Σ) , we have

$$(40) \quad |\psi(t)| \leq \left[1 - \frac{\varepsilon d^2}{(1 + \varepsilon)(1 + d^2)} \right] |\psi(t-1)| \quad \forall t > t_1,$$

which means that there exists $t_2 > t_1$ such that $|\psi(t_2)| < 1 + \varepsilon$. \square

4.3. Solutions in the “strict stability” set outside the ALIS.

THEOREM 5. *For any d , $0 < |d| \leq d^*$, with d^* , η given by (16), (17), we have*

(i) *Global attractiveness.* Any solution in $\{|\psi| \leq 1 - \eta\} \times \mathbb{R}$ exponentially approaches the ALIS. Moreover, a solution starting in $\{|\psi| \leq 1 - \eta\} \times \mathbb{R}$ remains in this set as long as it remains in the set $\{|x| \geq 1/(1 - \eta)\}$.

(ii) *Drift/Ejection.* If $|1/\alpha| > 1$ and for some time t_0 , a solution satisfies $|\psi(t_0)| \leq 1 - \eta$, then there exists a finite time t_1 such that $|\psi(t_1)| > 1 - \eta$. Hence, for $|1/\alpha| > 1$, there is no solution satisfying every $|\psi| \leq 1 - \eta$.

(iii) Moreover, while a solution remains in the set

$$\left\{ \frac{d^2(1 + \xi + n_0)n_1}{1 + d^2n_0^2} \leq \psi \leq 1 - \eta \right\} \times \{|x| \leq \xi\},$$

respectively, in the set

$$\left\{ -(1 - \eta) \leq \psi \leq -\frac{d^2(1 + \xi + n_0)n_1}{1 + d^2n_0^2} \right\} \times \{|x| \leq \xi\},$$

it crosses the ALIS at each time t (respectively, it remains on the same side).

Proof. (i) Since $|\psi| \leq 1 - \eta$ and $|x| \geq 1/(1 - \eta)$, then $|(\psi + d^2x)/(1 + d^2x^2)| \leq 1 - \eta$, a solution starting in $\{|\psi| \leq 1 + \eta\} \times \mathbb{R}$ remains in this set at least while it remains in the set $\{|x| \geq 1/(1 - \eta)\}$. To complete the proof of (i), with property (35), we only need to show that any solution remaining in $\{|\psi| \leq 1 - \eta\} \times \mathbb{R}$ enters the set $\{|x| < \xi\}$, (with ξ satisfying (33)) in a finite number of steps. In fact, let the constant $\xi' := (n_0 + \xi)/2$ be $n_0 < \xi' < \xi$ and $|x(t)| \geq \xi'$; then, from the equations of (Σ) and (18), we have

$$(41) \quad \left| \frac{x(t+1)}{x(t)} \right| \leq \left| \frac{1 + \alpha}{x(t)} \right| + (1 - \eta) < 1 - \eta + \frac{2n_0}{\xi + n_0} \eta < 1,$$

i.e., the absolute value of the x -component decreases exponentially as long as $|x| \geq \xi'$.

(ii) We have

$$(42) \quad \begin{aligned} \psi(t+1) = \psi(t) + & \frac{d^2G(\psi(t))(1 - \psi(t)G(\psi(t)))}{1 + d^2G(\psi(t))^2} \\ & + d^2[x(t) - G(\psi(t))] \frac{1 + \psi(t)(x(t) + G(\psi(t)))}{(1 + d^2G(\psi(t))^2)(1 + d^2x(t)^2)}. \end{aligned}$$

From (i), either the ψ -component of the solution leaves the interval $[-(1 - \eta), 1 - \eta]$ or, after a finite time, $x(t) - G(\psi(t))$ will be as small as we want. The result follows from a continuity argument and Theorem 3(c).

Statement (iii) is a direct consequence of Theorem 2(ii). \square

Remark. As in the discussion following Theorem 1, the global character of the attractiveness of this graph is a direct consequence of the use of a normalized algorithm. In general, nonnormalized algorithms, as treated by Praly (1990), may not lead to this kind of global result.

4.4. Solutions in the “critical stability” set.

THEOREM 6 (solutions in the sets B, D, F, H).

(i) As long as a solution remains in the set $\{(\psi, x)/1 - \eta < |\psi| < 1 + \varepsilon \text{ and } |x| > \chi\}$, the absolute value $|\psi|$ decays exponentially.

(ii) Any solution starting in the set $F \cup H$ (respectively, $B \cup D$) either enters the set G (respectively, C) or goes into the set E in a finite time.

Proof. Statement (i) follows exactly along the same lines as in (38)–(40).

(ii) From

$$\phi = \left(\frac{\psi x + d^2 x^2}{1 + d^2 x^2} \right) \frac{1}{x},$$

we easily obtain

$$\begin{aligned}(\psi, x) \in F &\Rightarrow \{\psi x > 1, x > 0\} \Rightarrow \phi > 1/x > 0, \\(\psi, x) \in H &\Rightarrow \{\psi x < -1, x < 0\} \Rightarrow \phi > 1/x \geq -(1 - \eta), \\(\psi, x) \in D &\Rightarrow \{\psi x > 1, x < 0\} \Rightarrow \phi < 1/x < 0, \\(\psi, x) \in B &\Rightarrow \{\psi x < -1, x > 0\} \Rightarrow \phi < 1/x < 1 - \eta,\end{aligned}$$

and the claim follows since from (i), $|\psi|$ decreases exponentially while $(\psi, x) \in B \cup D \cup F \cup H$. \square

THEOREM 7 (solutions in the set G). *For $|d|$ small enough, if $|1/\alpha|$ is larger than $1 + \varepsilon$, a 2-periodic orbit exists such that at least one of its points lies in the set*

$$G = \{(\psi, x)/1 - \eta < \psi < 1 + \varepsilon \text{ and } |x| < \chi\}.$$

Proof. According to Theorem A1 in the Appendix, for $|d|$ small enough, a 2-periodic orbit exists with its ψ component such that $1 - \psi = O(d^2)$. This implies that the 2-periodic orbit is contained in $F \cup G \cup H$ for a small enough d^2 . The result follows since from Theorem 6 the orbit cannot be entirely contained in $F \cup H$. \square

4.5. Boundedness of solutions.

THEOREM 8. *If $1/\alpha > -1$, all the solutions of (Σ) are bounded.*

Proof. With Theorem 4.1 of Egardt (1979) it is sufficient to prove that the sequence $\{|\psi(t)|\}$ is bounded for any solution of (Σ) . For this, we first show, from the second equation of (Σ) that when $|\psi| > |d|/2\sqrt{2}$, $|\psi(t+1)| > |\psi(t)|$ if and only if $\psi(t)x(t) \in (0, 1)$. The proof then follows by showing that, for $1/\alpha > -1$, there exists $\gamma > |d|/2\sqrt{2} > 0$ such that the points of the set $\Gamma \equiv \{(\psi, x)/|\psi| > \gamma; \psi x \in (0, 1)\}$ have no preimage in Γ (the trajectories starting in Γ leave it in one sampling time). This is combined with the relationship $|\psi(t+1)| < |\psi(t)| + |d|/2$ to show that if $\psi(0)$ satisfies

$$|\psi(0)| \leq \max \left\{ \frac{1 + 2 \max \{|\alpha|, |1 - \alpha|\}}{\max \{|\alpha|, |1 + \alpha|\}}, \frac{|d|}{2\sqrt{2}} \right\} + |d|,$$

then $\psi(t)$ satisfies the same inequality for all $t \geq 0$. If, on the other hand, $\psi(0)$ does not satisfy the above inequality, there exists a finite time T such that $\psi(T)$ does satisfy it (see details in España and Praly (1988)). \square

5. Global behavior of the solutions: Qualitative description and simulation results. Using the technical results of the previous sections (Theorems 1–8), we can explain the five stages of the solutions' behavior observed in simulation and mentioned in § 2. For this, use is made of the phase plane decomposition introduced at the beginning of § 4. Figures 1–5 are used to illustrate the system's dynamic behavior. The function h given by (4) has been plotted in the phase portrait part of each figure. As shown by Theorems 1(iii) and 2(iii), its graph, denoted by “ hg ,” approximates the RLIS and the ALIS, respectively, in their domain of definition.

5.1. The turbulent phase.

5.1.1. Explosive stage. According to Theorem 4(i), a solution in the sets A or I , either remains in the RLIS, which is the graph of a bounded function of ψ , or diverges exponentially from it (and, in practice, from its approximation (4)). This explains an exponential growth of the x -component, which becomes and remains large. Moreover, for a solution in the set I , at each time t , the x -component changes side with respect to

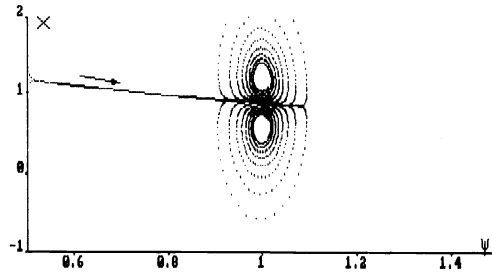


FIG. 1(a). Phase portrait, two stable focus of Σ^2 ($\alpha = 0.8$, $d^2 = 0.005$).

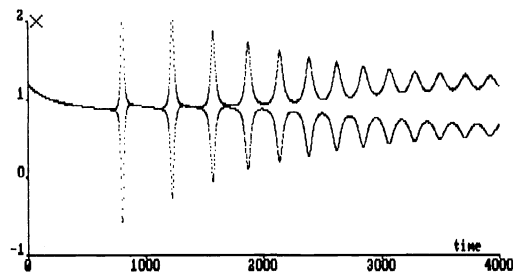


FIG. 1(b). Time response converging to a period-2 stable orbit ($\alpha = 0.8$, $d^2 = 0.005$).

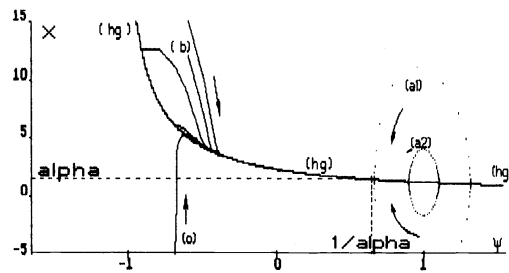


FIG. 2(a). Phase portrait with a stable node ($\alpha = 1.5$, $d^2 = 0.005$).

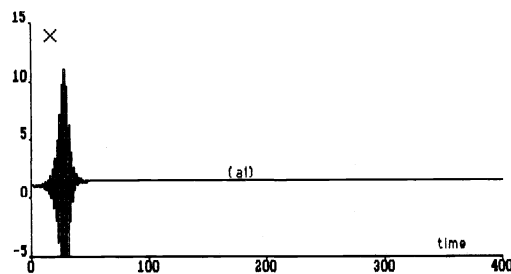


FIG. 2(b). Time response of solution (a1) ($\alpha = 1.5$, $d^2 = 0.005$).

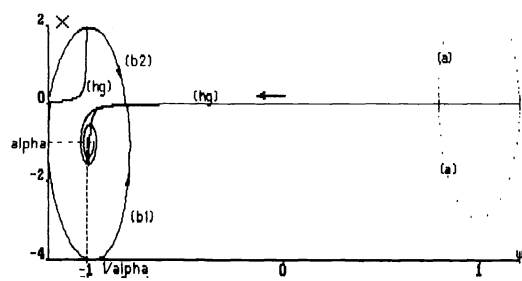


FIG. 3(a). Phase portrait with a stable focus ($\alpha = -1.008, d^2 = 0.005$).

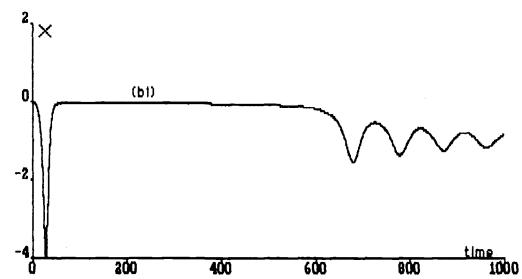


FIG. 3(b). Time response of solution (b1) ($\alpha = -1.008, d^2 = 0.005$).

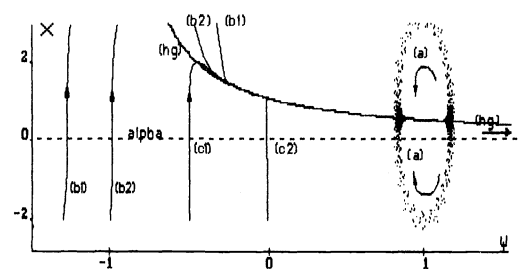


FIG. 4(a). Phase portrait with a saddle as equilibrium point ($\alpha = 0.1, d^2 = 0.005$).

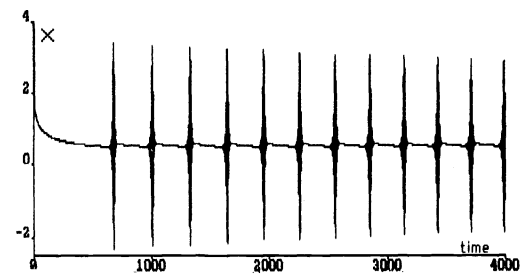
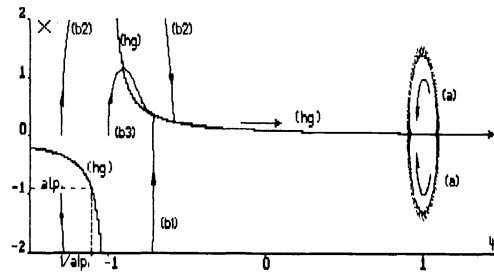
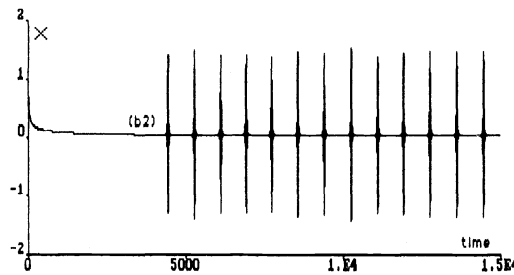


FIG. 4(b). Time response of solution (a) ($\alpha = 0.1, d^2 = 0.005$).

FIG. 5(a). Phase portrait with an unstable node ($\alpha = -0.9$, $d^2 = 0.005$).FIG. 5(b). Time response of solution (b2) ($\alpha = 0.9$, $d^2 = 0.005$).

this graph. This explains the “bursts” with very high frequency content on x (see solutions (a) in Figs. 2, 4, and 5). Conversely, for a solution in the set A , the x -component remains on the same side of the graph. It corresponds to a burst without oscillations (see solutions (b) in Figs. 3–5). We conclude that the explosive stage (a) takes place in the set A or I for any value of the disturbance and the reference (and therefore also in the ideal case).

5.1.2. Reinjection stage. Following Theorem 4(ii), a solution in the set A or I with a large x -component or in the set $B \cup D$ or $F \cup H$ has its ψ -component exponentially decaying. This occurs for any value of the disturbance and the reference and explains the reinjection of the solutions into the set E (see solutions (a) and (b) in Figs. 2–5).

5.1.3. Implosive stage. Theorem 5(i) states that, at least for a sufficiently small disturbance, as soon as a solution enters the set E , it is exponentially attracted toward the ALIS, which is the graph of a bounded function of ψ approximated by the set hg . This explains the exponential decrease of the x -component and, were it present, the fast decay of its high frequency content (see solutions (a) in Figs. 2, 4, and 5). Consequently, at least for small values of the disturbance, this stage occurs for any value of the reference and takes place in the set E .

5.2. The laminar phase (or the drift/ejection stage). Following Theorem 5(ii), when $|1/\alpha| > 1$, all the solutions leave the set E in a finite time. However, if before leaving the set E , they become close to the ALIS (in Figs. 1–5 one can see how the solutions practically converge to the graph (hg) approximating the ALIS), they finally leave that set “drifting” over the ALIS while its ψ -component grows with a speed of the order of d^2 (see also Theorem 3(c)). The solutions, very likely, enter the set G .

After entering the set G , a solution may either remain in it (see Fig. 1), go to the set I , thus restarting the explosive stage and possibly initiating the intermittent phenom-

enon, or go to the set $F \cup H$. In the latter case, Theorem 6 shows that the solution may either be reinjected into the set E , restarting the implosive stage, or returned to the set G . Intermittency may also take place in this case.

5.3. Possible 2-periodic orbits or limit cycles as ω -limits. According to Theorems A1 (in the Appendix) and 7, for a reference-to-disturbance ratio strictly smaller than 1 and for a disturbance sufficiently small, a 2-periodic orbit exists with at least one point in G and the other in $F \cup G \cup H$. Each point of this orbit is an attractive focus of Σ^2 if reference and disturbance have the same signs and a repellent focus in the case of opposite signs. In the former case, intermittency may disappear asymptotically while the solutions converge toward a 2-periodic orbit (see Fig. 1 and solution (a) in Fig. 4). When the reference and the disturbance have different signs, it is hypothesized that the solutions either exhibit a permanent intermittency (see Fig. 5) or, due to the occurrence of a supercritical Hopf bifurcation of the 2-periodic orbit, have an ω -limit comprised of two limit cycles of Σ^2 each surrounding a fixed point of Σ^2 .

5.4. High sensitivity with respect to the initial conditions. From simulations and the approximations given in Theorems 1(iii) and 2(iii), it seems that the ALIS and the RLIS are smoothly connected through the set G (see Figs. 4 and 5). From Remark 1 following Theorem 3(b), if the fixed point lies in the RLIS, the portion of this set defined for $\psi > 1 + \varepsilon$ is unique. Its intersection with the boundary $\psi = 1 + \varepsilon$ being transverse, we expect that it is uniquely extended by an ALIS inside the strict stability region. Using this conjecture as a working hypothesis, the more a solution approaches the ALIS while it is in E , the more its evolution will be similar to the solutions in the RLIS when entering the set I . However, according to Theorem 3(b) and the remarks that follow, for a reference-to-disturbance ratio strictly smaller than 1 in absolute value and negative, the solutions in the RLIS starting in $\psi \in (1 + \varepsilon, \infty)$ are unbounded (the same as those starting in $(-\infty, 1/\alpha)$). On the other hand, the bigger its ψ -component is, the more the x -component of a solution in the set I , but not in the RLIS, is “pushed-away” (exponentially) from this invariant set. This reasoning shows the possibility of a very high sensitivity to initial conditions of solutions starting near the ALIS or, with Theorem 2, close to the graph of the function hg given in (4).

5.5. The desired behavior. Theorem A1 (see the Appendix) shows that, for a sufficiently small disturbance and a reference-to-disturbance ratio strictly larger than 1, the fixed point is exponentially stable and there is no other periodic solution. On the other hand, and under the same conditions, with Theorem 8, each solution remains in a compact set. This suggests that the fixed point is a global attractor. In this case, intermittency should not take place and the desired working conditions should be attained (see Figs. 2 and 3). Qualitatively speaking, this case most resembles the ideal case.

Summarizing, according to the reference-to-disturbance ratio α , three qualitatively different behaviors of the solutions of (Σ) can be predicted:

1. $|\alpha| > 1$ (high level excitation): bounded solutions, no intermittency, no periodic solution, a global attractive fixed point is conjectured, behavior similar to the ideal case.

2. $0 < \alpha < 1$ (low level excitation): bounded solutions, stable periodic solutions exist and are conjectured to be global attractors, the fixed point is a saddle, intermittency may occur but is conjectured to gradually disappear while converging asymptotically to a 2-periodic solution.

3. $-1 < \alpha \leq 0$ (low level excitation): unbounded solutions exist, unstable 2-periodic solutions exist, intermittency and/or possible nonlinear oscillations are present, the fixed point is an unstable node.

Since α is a relative quantity, drastic qualitative changes of the system's behavior may be expected when both r and d are close to zero, which is the natural working condition for an adaptive linear regulator.

6. A means to prevent intermittency: The dead zone. We study here the effects of an empirical modification to the second equation of (Σ) (and (Σ_1)). For some $\delta > 0$, we call the set $D_\delta = \{x / |x - \alpha| < \delta\}$ the δ -dead zone and substitute μ in (Σ) by:

$$\mu = \begin{cases} 0 & \text{iff } x \in D_\delta, \\ 1 & \text{iff } x \notin D_\delta. \end{cases}$$

It is expected that this modification will interrupt the drift stage of the solutions near the ALIS when $|1/\alpha| > 1$. To examine the validity of this, let us first introduce the following definition:

$$\nu(\eta, d^2) := \frac{d^2 n_0 (n_0 + n_1) (n_0 + 1)}{(1 - \lambda(\eta, d^2))},$$

where η and λ are given respectively by (17) and (30). We now make the following assertion.

Assertion. (i) If $|1/\alpha| > 1$ and d and δ are such that

$$\delta < \frac{1 - \alpha + \eta\alpha}{2 - \eta} - \nu(\eta, d^2),$$

there is no solution of (Σ) with δ -dead zone satisfying for all t : $|\psi(t)| < 1 - \eta$.

(ii) There is no solution leaving the set $\{|\psi| \leq 1\}$ if and only if $\delta \geq 1 + |\alpha|$.

Proof. (i) From the exponential attractiveness property of the ALIS and its approximation by the graph of h given by (4), (see Theorem 3 and (36)), we see that if $\alpha + \delta < h(1 - \eta) - \nu(\eta, d^2)$, any solution remaining in $\{|\psi| < 1 - \eta\}$ enters a set (the band around h of radius $\nu(\eta, d^2)$) whose intersection with the dead zone is empty for $|\psi| < 1 - \eta$. In this set, then, $\mu = 1$, the ψ -component is strictly increasing (Theorem 4(c) and (37)), and the solutions necessarily leave the domain $\{|\psi| < 1 - \eta\}$ in a finite time.

(ii) From the definition of (Σ) when $\mu = 1$ we have the following implications:

$$\begin{aligned} \{|\psi(t)| < 1 \text{ and } \psi(t+1) > 1\} &\Leftrightarrow \{1 > \psi(t) > 1 - d^2 x(1 - x)\}, \\ \{|\psi(t)| < 1 \text{ and } \psi(t+1) < -1\} &\Leftrightarrow \{-1 < \psi(t) < -1 - d^2 x(1 + x)\}. \end{aligned}$$

Thus, the solutions of (Σ) with dead zone will not leave the domain $\{|\psi| \leq 1\}$ if and only if the right hand side of both implications are realized inside the dead zone only. But this is clearly the case when $\delta > |\alpha| + 1$. \square

Remarks. The dead-zone modification may fail to work if an upper limit of the disturbance is not known. When it works, i.e., when δ is bigger than, say, some $\delta^* > (1 - \alpha + \eta\alpha)/(2 - \eta) + \nu(\eta, d^2)$, the original ALIS is transformed into a new one (possibly not given by the graph of a continuous function any more) containing a portion of the graph of h . Since h is a monotonically decreasing function in $|\psi| < 1 - \eta$, the part of the graph of h coinciding with the new ALIS is confined to the right of the interval $|\psi| < 1 - \eta$. However, the movement of the solutions over the ALIS is also in the sense of growing ψ 's. Consequently, in this case the modified scheme will very likely stop the solutions' drift near the ALIS before they enter into the "instability domain."

7. Concluding remarks. The ALIS and the RLIS play a key role in the qualitative and geometrical description of the phase portrait of our example. It can be shown that these two sets exist in more general systems since their definitions rest on some general properties of the adaptive systems (see, for instance, Praly (1990)). We can thus state that the intermittent behavior is the result of the absence of a global attractor in the ALIS combined with a property of the algorithm of maintaining all the signals bounded despite a model mismatch (L^∞ -robustness). In fact, since the ALIS is normally defined in a bounded open set of the parameter space (in general, the set of parameters mapped into the open unitary circle by the eigenvalues of the regressor model), the lack of a global attractor in it implies that some trajectories approaching the ALIS will eventually leave the strict stability set of parameters. This may be a very slow “quasi-stable” process. On the other hand, the L^∞ -robustness is responsible for the “reinjection mechanism” into the domain of attraction of the ALIS, provoking the abandon of the turbulent phase and the restart of the cycle. Since this reinjection is guaranteed by the (desirable) robustness, we can thus say that intermittency is essentially conditioned by the dynamics in the ALIS and thus, that any palliative to this phenomenon passes by an “adequate” modification of the dynamics in this set. However, the dynamics in the ALIS (and the ALIS itself) depend on the exogenous signals. Consequently, for any “good” modification of the ALIS dynamics it may be possible to find a “counter-example” given by a particular combination of model mismatch and reference signals. Moreover, the modifications introduced (to the algorithm and to the ALIS) may even exacerbate the situation. For instance, Rey, Bitmead, and Johnson (1989), reported that a previous (“unhelped”) nonintermittent system may become intermittent after the addition of leakage. This possibility seems less likely when a dead zone is used since its working principle consists in transforming part of the ALIS into the corresponding set of attractive bounded solutions of the “frozen system.” When the parameters are frozen, the drift phase is necessarily eliminated at least while the dead zone is active.

Clearly, more research must be done to find algorithms assuring a robust global attractor in the ALIS for (at least) a specified family of disturbances or model mismatches with a practical meaning. A possible general approach could be to stop the adaptation when some “ad-hoc” mechanism detects the drift phase. This decision can be taken, for instance, when the calculated increment in the parameters (possibly averaged) is smaller than a prespecified threshold. Note that here is not the error that counts, as in the dead zone, but its correlation with the regressor vector. This correlation should be viewed as an approximation of the gradient of the mean value of the error with respect to the parameters. When the reference signal is sufficiently persistently exciting or, more precisely, using a concept coined by Ioannou and Kokotovic (1983), persistently dominantly exciting, the ALIS has a natural global attractor in it; it corresponds to the desired working conditions. This attractor is hyperbolic (see Anderson et al. (1986)) and thus, structurally robust. However, for large disturbances (or model-mismatch), this attractor may cease to be globally attractive, disappear, or be pushed out of the ALIS in the RLIS region; this last circumstance motivates a bifurcation analysis. Any of these conditions can be at the origin of an intermittent behavior with persistent excitation, but in particular, when the attractor leaves the ALIS, the desired working conditions can never be attained even if by some means intermittency could ever be avoided. Clearly we can “solve the problem” by adding excitation in the dominant frequency range to the reference signals, but this does not imply that the original objective will be satisfied. On the other hand, when there is no persistent excitation (or there is but with a very low energy level or a bad frequency content), the functioning regime becomes very critical since the desired working conditions may not correspond to an hyperbolically stable set. In this case, a

very weak robustness of many properties is to be expected. Indeed, very different qualitative behaviors can be close to one another and the system may easily switch from the desired working condition to an intermittent behavior or a self-oscillating mode. With respect to the last mode of operation, we see, for the example considered, that a 2-periodic orbit (self-oscillating model) appears as a result of the bifurcation provoked by the expulsion of the desired working conditions out of the ALIS. This 2-periodic oscillation has also been encountered by other authors in a similar example (Sethares and Mareels (1991), Rey, Bitmead, and Johnson (1991)). Actually, this is a particularity of the example (one-dimensional parameter space) and of the excitation condition (constant reference and disturbance signals). In general, these self-oscillating modes, which determine the frequency content of the “bursts,” may have any period or not exist at all (see España, (1990), (1991)). When the ALIS has no attractor in it and there are no self-oscillations or there are but they are not attractive (see $1/\alpha < -1$ in our example), a nonperiodic permanent intermittent regime is likely to take place. This is favored by the fact that, while reentering in the “stability-domain,” the trajectories are attracted toward a region (near the ALIS) where the solutions are highly sensitive with respect to the initial conditions. As pointed out by Bergé, Pomeau, and Vidal (1984), this, together with the absence of a periodic or quasi periodic attractive solution, seems to be at the origin of strange attractors.

Summarizing, we may conjecture that the antidotes for intermittency, like leakage or internal model principle, whose active principle is not based in stopping the solutions in its drift phase near the ALIS (they just modify the dynamics in it), could provoke the undesired effect if combined with a particular excitation and/or model mismatch. The schemes based on freezing the parameters upon detecting the drift phase (the dead-zone approach included) may also need a priori information of the disturbances to succeed, but, even if they can possibly fail to do the job, they are less likely to provoke the undesired effect. They are based on a characteristic of the phenomenon that is independent of the excitation conditions. Nevertheless, none of them solve the problem of the attractive self-oscillations. Strongly dominantly exciting references signals produce robust desired working conditions. Otherwise, if no precautions are taken, the system may easily switch among very different qualitative behaviors.

Appendix. Local behavior near the equilibrium point and period-2 solutions. The analysis of the local behavior of (Σ) near its equilibrium $(\psi, x) = (1/\alpha, \alpha)$ is based in the Jacobian matrix of (Σ) at this point:

$$J = \begin{pmatrix} -1/\alpha & -\alpha \\ -d^2/(1 + d^2\alpha^2) & 1/(1 + d^2\alpha^2) \end{pmatrix}.$$

The product of its eigenvalues is $P = -1/\alpha$ and their sum is given by

$$(A) \quad S = P + 1 - \frac{d^2}{P^2 + d^2}.$$

In terms of S and P , exponential stability of the equilibrium point is given by

$$(B) \quad 1 - S + P > 0$$

$$(C) \quad 1 + S + P > 0,$$

$$(D) \quad 1 - P > 0,$$

and the eigenvalues are real if

$$(E) \quad S^2 - 4P \geq 0.$$

Thus, the equilibrium point is exponentially stable if and only if $1/\alpha \in (-1, p)$, where $p \in (\frac{1}{2}, 1)$ is the unique solution of $2(1-p)(p^2 + d^2) = d^2$. Consequently, unless $1/\alpha \in (-1, p)$, the “desired working conditions” do not correspond to a stable equilibrium point. The points of intersection of curves (A) and (E) correspond to the transition from real to complex eigenvalues and vice versa. They all occur for α negative, and we may have a stable ($\alpha < -1$) or an unstable ($\alpha > -1$) equilibrium point and a node or a focus depending on d^2 .

Independently of the value of d^2 , for α near -1 , the equilibrium point passes from being an attractive focus ($\alpha < -1$) to a repellent one ($\alpha > -1$) verifying the conditions for a postcritical Hopf bifurcation (see Iooss (1979)). For the particular value $\alpha = -1$ the whole ψ -axis is a set of fixed points implying a global bifurcation.

For α positive, we have either a stable node ($1/\alpha < p$) or a saddle ($1/\alpha > p$). When $1/\alpha$ crosses the value p , an eigenvalue passes through -1 , and a stable period-2 solution bifurcates from the stable fixed point while the latter becomes unstable (see Arnold (1983) or Iooss (1979)). The 2-periodic solutions can be determined evaluating the roots of the equation $\Sigma^2(\psi, x) - (\psi, x) = 0$, or, equivalently, for d nonzero, by computing the solution of

$$(A.1) \quad \begin{aligned} F_x(x, \psi, d) &:= -(\psi + d^2\phi)(-\psi x + 1 + \alpha) + (1 + \alpha - x) = 0 \\ F_\psi(x, \psi, d) &:= \phi + \frac{(-\psi x + 1 + \alpha)[1 - (\psi + d^2\phi)(-\psi x + 1 + \alpha)]}{1 + d^2(-\psi x + 1 + \alpha)^2} = 0, \end{aligned}$$

where $\phi = x(1 - \psi x)/(1 + d^2 x^2)$. For $d = 0$, the above system has three solutions:

$$(A.2) \quad (\psi_0 = \alpha^{-1}, x_0 = \alpha), \quad \left(\psi_{1,2} = 1, x_{1,2} = \frac{1 + \alpha \pm \sqrt{1 - \alpha^2}}{2} \right).$$

The first one is the equilibrium point; the two others exist if and only if the disturbance-to-reference ratio $1/\alpha$ is larger (in modulus) than 1. We make the following observation.

LEMMA (Existence of periodic solutions; Pomet, Coron, and Praly (1990)). *A necessary condition for $(\psi_{\text{per}}(t, d), x_{\text{per}}(t, d))$ to be a period- T solution of (Σ) that remains bounded as d goes to 0 is that the accumulation point of its initial condition be one of the three points in (A.2).*

To show that the existence of zeros $(\psi_{1,2}, x_{1,2})$ of (A.1) given by (A.2) is also sufficient for having period-2 solutions, use is made of the implicit function theorem with the following expression of the Jacobian matrix of (A.1) (nonsingular for $|\alpha| < 1$):

$$\partial F(x_{1,2}; \psi_{1,2}; 0) = \begin{bmatrix} 0 & \mp \sqrt{1 - \alpha^2} \\ \mp \sqrt{1 - \alpha^2} & * \end{bmatrix},$$

where the $(2, 2)$ -term is unimportant. Compared with our discussion on the stability of the fixed point of (Σ) we note that for $|d|$ small enough, the period-2 solutions emerge not only when an eigenvalue of J passes through -1 ($\alpha \approx +1$), from the stability side (for period-2 bifurcation condition), but also when a pair of conjugate eigenvalues of J crosses the unit circle at $\alpha = -1$. The latter corresponds to a global bifurcation. To summarize, we have the following theorem.

THEOREM A1 (Critical elements, España and Praly (1988)). (i) *The system (Σ) has a unique fixed point for all α different from 0 or -1 . It is the solution corresponding to the control objective. It is exponentially stable for $1/\alpha \in (-1, p)$ and exponentially unstable for $1/\alpha \notin [-1, p]$.* (ii) *For any $|\alpha| < 1$ we can find a strictly positive constant*

d_0 such that if $|d| \leq d_0$, there exists two locally unique period-2 solutions that can be approximated by

$$(A.3) \quad \psi_{1,2} = 1 - \frac{\alpha d^2}{2} \frac{1 + \alpha \mp \sqrt{1 - \alpha^2}}{2} + O(d^4), \quad x_{1,2} = \frac{1 + \alpha \pm \sqrt{1 - \alpha^2}}{2} + O(d^2).$$

These solutions are foci of Σ^2 , exponentially stable for $\alpha > 0$, exponentially unstable for $\alpha < 0$, and with a pseudoperiod approximated by $T = 2\pi/d(2(1 - \alpha^2)^{1/2})$.

Acknowledgments. The authors thank the reviewers for their thorough work. Their excellent comments and pertinent suggestions contributed immeasurably to improving the quality of this article.

REFERENCES

- B. D. O. ANDERSON (1985), *Adaptive systems, lack of persistence of excitation and bursting phenomena*, Automatica, 21, pp. 247–258.
- B. D. O. ANDERSON, R. R. BITMEAD, C. R. JOHNSON, P. V. KOKOTOVIC, R. L. KOSUT, I. M. Y. MAREELS, L. PRALY, AND B. D. RIEDLE (1986), *Stability of Adaptive Systems: Passivity and Averaging Analysis*, MIT Press, Cambridge, MA.
- V. I. ARNOLD (1983), *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, Berlin.
- A. BENVENISTE, M. METIVIER, AND P. PRIOURET (1987), *Algorithmes Adaptatifs et Approximations Stochastiques: Théorie et Applications*, Masson, Paris.
- M. BODSON, S. SASTRY, B. D. O. ANDERSON, M. Y. MAREELS, AND R. R. BITMEAD (1986), *Nonlinear averaging theorems, and the determination of parameter convergence rates in adaptive control*, Systems Control Lett., 7, pp. 145–157.
- P. BERGE, Y. POMEAU, AND CH. VIDAL (1984), *L'Ordre dans le Chaos*, Hermann, Paris.
- B. EGARDT (1979), *Stability of Adaptive Controllers*, Springer-Verlag, New York, Berlin.
- H. ELLIOT AND G. C. GOODWIN (1984), *Adaptive implementation of the internal model principle*, in Proceedings of the 23rd IEEE Conference on Decision and Control, Las Vegas, NV.
- M. ESPAÑA (1991), *Intermittent phenomena in adaptive systems: A case study*, Automatica., 27, pp. 717–720.
- (1991), *Intermittency and self-oscillations in adaptive systems*, in Proceedings, VI IFAC Symposium on Automation of Mining, Mineral, and Metal Processing, Buenos Aires, Argentina, October.
- M. ESPAÑA AND L. PRALY (1988), *On the Global Dynamics of Adaptive Systems*, Internal Report CAI, Fontainebleau, France.
- G. C. GOODWIN AND K. S. SIN (1984), *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ.
- M. P. GOLDEN AND B. E. YDSTIE (1988), *Bifurcations in model reference adaptive control systems*, Systems Control Lett., 11, pp. 413–430.
- J. K. HALE (1980), *Ordinary Differential Equations*, Krieger Publishing Company, Huntington, NY.
- P. A. IOANNOU AND P. V. KOKOTOVIC (1983), *Adaptive Systems with Reduced Models*, Lecture Notes in Control and Inform. Sci., Vol. 47, Springer-Verlag, New York, Berlin.
- G. IOOSS (1979), *Bifurcation of Maps and Applications*, North-Holland, Amsterdam.
- M. JAÏDANE-SAÏDANE AND O. MACCHI (1988), *Quasi-periodic self-stabilization of adaptive ARMA predictors*, Internat. J. Adaptive Control Signal Processing, March.
- L. LJUNG (1977), *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, August.
- L. LJUNG AND T. SÖDERSTRÖM (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA.
- K. S. NARENDRA AND A. ANNASWAMY (1986), *Robust adaptive control in the presence of bounded disturbances*, IEEE Trans. Automat. Control, April, pp. 306–315.
- Y. POMEAU AND P. MANVILLE (1980), *Intermittent transition to turbulence in dissipative dynamical systems*, Comm. Math. Phys., 74, pp. 189–197.
- J.-B. POMET, J. M. CORON, AND L. PRALY (1990), *On the periodic solutions of adaptive systems in the presence of periodic forcing terms*, Math. Control Signals Systems, 3, pp. 373–399.

- L. PRALY (1985), *A geometric approach for the local analysis of a one-step-ahead adaptive controller*, in Proc. of the 4th Yale Workshop on Applications of Adaptive Systems Theory, New Haven, CT, June.
- (1988), *Oscillatory behavior and fixes in adaptive linear control: a worked example*, in Proc. of the IFAC Workshop on Robust Adaptive Control, Newcastle, Australia, August.
- (1990), *Topological orbital equivalence with asymptotic phase for a two time-scales discrete-time system*, Math. Control Signals Systems, 3, pp. 225–253.
- G. J. REY, R. BITMEAD, AND C. R. JOHNSON (1991), *The dynamics of bursting in simple adaptive feedback systems with leakage*, IEEE Trans. Circuits and Systems, 38, pp. 426–488.
- B. D. RIEDLE AND P. V. KOKOTOVIC (1986), *Integral manifold of slow adaptation*, IEEE Trans. Automat. Control, April.
- M. SHUB (1987), *Global Stability of Dynamical Systems*, Springer-Verlag, New York, Berlin.
- W. A. SETHARES AND M. Y. MAREELS (1991), *Dynamics of an adaptive hybrid*, IEEE Trans. Circuits and Systems, 38, pp. 1–11.

OPTIMAL CONTROL FOR INTEGRODIFFERENTIAL EQUATIONS OF PARABOLIC TYPE*

GIUSEPPE DA PRATO[†] AND AKIRA ICHIKAWA[‡]

Abstract. Quadratic control problems for integrodifferential equations of parabolic type are considered. A state-space representation of the system is obtained by choosing an appropriate product space. By using the standard method based on Riccati equation, a unique optimal control over a finite horizon and under a stabilizability condition is obtained and the quadratic problem over an infinite horizon is solved. It is shown that the approach is also valid for some integrodifferential equations of different types. Two examples covered by the model are given.

Key words. optimal control, stabilizability, integrodifferential equations

AMS subject classifications. 93D15, 93C22, 93C25

1. Introduction. Let H and U be Hilbert spaces. Consider the control system

$$(1) \quad \begin{cases} y'(t) = Ay(t) + \int_0^t K(t-r)y(r) dr + Bu(t), \\ y(0) = y_0, \end{cases}$$

where A is the infinitesimal generator of an analytic semigroup e^{tA} in H . We denote by $D(A)$ the domain of A and by $|\cdot|_{D(A)}$ the graph norm of A . $K(\cdot)$ is an $L(D(A); H)$ -valued operator, and $B \in L(U; H)$. Under suitable conditions (see Hypothesis 1 below) there exists a resolvent operator (see [3], [13], [16]) associated with (1) and a unique classical solution to (1). For each $u \in L^2(0, T; U)$ we can define a mild solution to (1) in $C([0, T]; H)$. We then wish to minimize the functional

$$(2) \quad J(u) = \int_0^T \{ |My(t)|^2 + |u(t)|^2 \} dt + \langle Gy(T), y(T) \rangle$$

over all $u \in L^2(0, T; U)$. Here $M \in L(H; H_0)$, H_0 is a Hilbert space, and $G \in L^+(H)$ is the space of selfadjoint nonnegative operators on H . Under a stabilizability condition (see Hypothesis 4 below) we also wish to minimize the functional

$$(3) \quad J(u) = \int_0^\infty \{ |My(t)|^2 + |u(t)|^2 \} dt$$

over all $u \in L^2(0, \infty; U)$. To our knowledge there is no direct method to solve these problems. In this paper we give a state-space representation of (1) similar to those in [15] and [19]. As in [9]–[11] we then reduce our problems to linear quadratic problems of standard type [1].

We recall some fundamental results concerning the resolvent operator associated with (1). It is convenient to introduce equations

$$(4) \quad \begin{cases} y'(t) = Ay(t) + \int_0^t K(t-r)y(r) dr, \\ y(0) = y_0, \end{cases}$$

* Received by the editors October 22, 1990; accepted for publication (in revised form) January 17, 1992.

[†] Scuola Normale Superiore di Pisa, Piazza dei Cavalieri, 7, 56100, Pisa, Italy.

[‡] Department of Electrical Engineering, Shizuoka University, Hamamatsu 432, Japan.

$$(5) \quad \begin{cases} y'(t) = Ay(t) + \int_0^t K(t-r)y(r) dr + f(t), \\ y(0) = y_0, \end{cases}$$

In [6] and [16] the existence of a resolvent operator for (4) is shown under the following conditions.

Hypothesis 1.

- (i) $K(\cdot) \in L^1(0, \infty; L(D(A); H))$.
- (ii) For all $h \in D(A)$, the Laplace transform $\tilde{K}(\cdot)h$ can be extended to a sector $S = \{\lambda \in \mathbb{C} : \lambda \neq \omega, |\arg(\lambda - \omega)| < \varphi\}$, where $\omega \in \mathbb{R}, \varphi \in]\pi/2, \pi[$.
- (iii) There exist $\beta \in]0, 1]$ and $c > 0$ such that $|\lambda^\beta \tilde{K}(\cdot)h| \leq c|h|_{D(A)}$, $\lambda \in S$, $h \in D(A)$.

The following result is proved in [3] and [16].

THEOREM 1.1. *There exists an analytic resolvent operator $R(t) \in L(H; D(A))$, $t \geq 0$, such that*

- (i) $R(t)y_0$ is continuous for any $y_0 \in H$ and $R(0) = I$.
- (ii) For each $y_0 \in D(A)$ and $T > 0$,

$$R(t)y_0 \in C([0, T]; D(A)) \cap C^1([0, T]; H)$$

and it satisfies (4).

- (iii) For each $y_0 \in D(A)$ and $f \in C^\alpha([0, T]; H)$ (α -Hölder continuous), $y(t)$, given by

$$y(t) = R(t)y_0 + \int_0^t R(t-r)f(r) dr,$$

is a unique classical solution (see [3]), in

$$C([0, T]; H) \cap C([0, T]; D(A)) \cap C^1([0, T]; H).$$

- (iv) There exist $r_0 > 0$ and $\varphi_0 \in]\frac{\pi}{2}, \varphi]$ such that for any $\lambda \in S$ with $|\lambda| \geq r_0$, $|\arg \lambda| \leq \varphi_0$, the linear operator $\lambda - A - \tilde{K}(\lambda) : D(A) \rightarrow H$ is invertible and $(\lambda - A - \tilde{K}(\lambda))^{-1} \in L(H; D(A))$ coincides with the Laplace transform of $R(t)$.

For each $y_0 \in H$ and $u \in L^2(0, T; U)$

$$(6) \quad y(t) = R(t)y_0 + \int_0^t R(t-r)Bu(r) dr$$

is well defined and is in $C([0, T]; H)$. It is a mild solution of (1) in the sense

$$(7) \quad y(t) = e^{tA}y_0 + \int_0^t e^{(t-r)A} \int_0^r K(r-s)y(s) ds dr + \int_0^t e^{(t-r)A} Bu(r) dr.$$

Note that the cost function (2) makes sense for the mild solution.

For later use we establish additional properties of $R(t)$ that are not given in [3]. Let $D_A(\alpha, 2)$, $\alpha \in]0, 1[$ be the real interpolation space between $D(A)$ and H . Consider the problem

$$(8) \quad y'(t) = Ay(t) + f(t), \quad y(0) = y_0.$$

THEOREM 1.2. (i) Let $y_0 \in D_A(\frac{1}{2}, 2)$, and let $f \in L^2(0, T; H)$. Then the mild solution of (8) lies in

$$L^2(0, T; D(A)) \cap W^{1,2}(0, T; H) \subset C([0, T]; D_A(\frac{1}{2}, 2)).$$

There exists a unique solution y to (4) in $L^2(0, T; D(A)) \cap W^{1,2}(0, T; H)$. Hence $R(t)y_0, R \star f \in L^2(0, T; D(A)) \cap W^{1,2}(0, T; H)$.

(ii) Let $y_0 \in D(A)$, $f \in W^{1,2}(0, T; H)$, and $Ay_0 + f(0) \in D_A(\frac{1}{2}, 2)$. Then the mild solution of (8) lies in

$$W^{1,2}(0, T; D(A)) \cap W^{2,2}(0, T; H) \subset C^1([0, T]; D_A(\frac{1}{2}, 2)).$$

Moreover, there exists a unique solution y to (4) in $W^{1,2}(0, T; D(A)) \cap W^{2,2}(0, T; H)$. Hence $R(t)y_0, R \star f \in W^{1,2}(0, T; D(A)) \cap W^{2,2}(0, T; H)$.

Proof. The first assertion in (i) is well known [17]. To show the second assertion of (i) we consider the corresponding integral equation of the type (7). For a small T we apply a contraction-mapping theorem on $L^2(0, T; D(A))$. The general case then follows by splitting the interval into small subintervals. The first assertion in (ii) is proved as in [15]. The second part of (ii) follows by raising regularity and considering a contraction mapping on $W^{1,2}(0, T; D(A))$. See [15] for details. \square

To give a state-space representation [8], [9] of (1) we consider

$$(9) \quad \begin{cases} y'(t) = Ay(t) + \int_{-\infty}^t K(t-r)y(r) dr, \\ y(0) = y_0, \\ y(\theta) = y_1(\theta), \theta \in]-\infty, 0[, y_1 \in L^2(-\infty, 0; D(A)). \end{cases}$$

We now rewrite this as

$$(10) \quad \begin{cases} y'(t) = Ay(t) + \int_0^t K(t-r)y(r) dr + f(t), \\ y(0) = y_0, \end{cases}$$

where $f(t) = \int_{-\infty}^0 K(t-\theta)y_1(\theta) d\theta \in L^2(0, \infty; H)$.

Hypothesis 2. $K(\cdot) \in L^2(0, \infty; L(D(A); H))$.

If we assume Hypothesis 2 is true, then the operator \mathcal{K} defined by

$$(11) \quad \mathcal{K}y_1 = \int_{-\infty}^0 K(-\theta)y_1(\theta)d\theta, \quad y_1 \in L^2(-\infty, 0; D(A))$$

lies in $L(L^2(-\infty, 0; D(A)), H)$. Moreover, $f \in W^{1,2}(0, T; H)$ for any $y_1 \in W^{1,2}(-\infty, 0; D(A))$ since

$$f'(t) = K(t)y_1(0) + \int_{-\infty}^0 K(t-\theta)y_1'(\theta) d\theta \in L^2(0, T; H).$$

Note also $f(0) = \mathcal{K}y_1$. Using these observations and Theorem 1.2, we have the following corollary.

COROLLARY 1.3. (i) For each $y_0 \in D_A(\frac{1}{2}; 2)$ and $y_1 \in L^2(-\infty, 0; D(A))$ there exists a unique solution to (10) (and hence to (9)) in the space $L^2(0, T; D(A)) \cap W^{1,2}(0, T; H)$.

(ii) Assume Hypothesis 2, and let $y_1 \in W^{1,2}(-\infty, 0; D(A))$, $y_1(0) = y_0$, and $Ay_0 + \mathcal{K}y_1 \in D_A(\frac{1}{2}; 2)$. Then there exists a unique solution to (9) in

$$W^{1,2}(-\infty, 0; D(A)) \cap W^{2,2}(0, T; H) \subset C^1([0, T]; D_A(\frac{1}{2}; 2)).$$

Proof. Part (i) follows directly from Theorem 1.2(i). Under Hypothesis 2 $f \in W^{1,2}(0, T; H)$ and the assumptions in (ii) of Theorem 1.2 are satisfied. Hence (10) has a unique solution in $W^{1,2}(0, T; D(A)) \cap W^{2,2}(0, T; H)$. Since $y_1(0) = y_0 \in D(A)$, there exists a unique solution to (5) in $W^{1,2}(-\infty, T; D(A))$. \square

Now we write (9) in the form

$$\begin{aligned} y'(t) &= Ay(t) + \int_{-\infty}^t K(-\theta)y(t+\theta) d\theta, \\ y(0) &= y_0, \\ y(\theta) &= y_1(\theta), \quad \theta \in]-\infty, 0[. \end{aligned} \tag{12}$$

This is a delay equation with infinite delay. A more general delay equation, but with finite delay, was considered in [15], and a semigroup was constructed on the product space $D_A(\frac{1}{2}; 2) \times L^2(-r, 0; D(A))$. To obtain a similar result we assume Hypothesis 2 and rewrite (12) as

$$\begin{aligned} y'(t) &= Ay(t) + \mathcal{K}y_t, \\ y(0) &= y_0, \\ y(\theta) &= y_1(\theta), \quad \theta \in]-\infty, 0[. \end{aligned} \tag{13}$$

where $y_t(\cdot) = y(t + \cdot)$. The following result is a modification of [15, Thms. 4.1 and 4.2] for the case with infinite delay.

THEOREM 1.4. Assume Hypotheses 1 and 2 are true, and let y be the unique solution of (12) for $y_0 \in D_A(\frac{1}{2}; 2)$ and $y_1 \in L^2(-\infty, 0; D(A))$, which lies in $L^2(0, T; D(A)) \cap W^{1,2}(0, T; H)$ for any $T > 0$. Then the map

$$\underline{S}(t) : \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \rightarrow \begin{pmatrix} y(t) \\ y_t(\cdot) \end{pmatrix} \tag{14}$$

on $\underline{Z} = D_A(\frac{1}{2}; 2) \times L^2(-\infty, 0; D(A))$ is a strongly continuous semigroup. Its infinitesimal generator is given by

$$\mathcal{A} \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} = \begin{pmatrix} Ay_0 + \mathcal{K}y_1 \\ \frac{dy_1}{d\theta} \end{pmatrix}, \tag{15}$$

$$D(\mathcal{A}) = \{(y_0, y_1) \in \underline{Z} : y_1 \in W^{1,2}(-\infty, 0; D(A)), \\ y_1(0) = y_0, Ay_0 + \mathcal{K}y_1 \in D_A(\frac{1}{2}; 2)\} \tag{16}$$

Proof. The only difference between (13) and the equation in [15] is the length of the memory involved. Hence one could repeat the proofs in [15]. However, we shall give a different proof for the characterization of the generator. Note first that the strong continuity and the semigroup property of $\underline{S}(t)$ follow from Corollary 1.3(i). We now show that \mathcal{A} , given by (15) and (16) is the infinitesimal generator of the semigroup $\underline{S}(t)$. Choose $[y_0, y_1]' \in D(\mathcal{A})$; then $\underline{S}(t)[y_0, y_1]' = [y(t), y_t(\cdot)]$, where $y(t)$ is the solution of (13) and hence of (9). Then by Corollary 1.3 (ii) we have

$$\lim_{t \rightarrow 0} \frac{y(t) - y_0}{t} = y'(0) = Ay_0 + \mathcal{K}y_1 \quad \text{in } D_A\left(\frac{1}{2}; 2\right),$$

$$\lim_{t \rightarrow 0} \frac{y_t(\cdot) - y_1(\cdot)}{t} = \frac{dy_1}{d\theta} \quad \text{in } L^2(-\infty, 0; D(A)).$$

This implies that the infinitesimal generator of the semigroup $\underline{S}(t)$ coincides with \mathcal{A} on $D(\mathcal{A})$ and is an extension of \mathcal{A} . To see that \mathcal{A} is in fact the generator we need to show only that the resolvent set of \mathcal{A} is nonempty. Now choose $\lambda > 0$ and consider

$$(\lambda - \mathcal{A})[y_0, y_1]' = [z_0, z_1]' \in \underline{Z}.$$

This is equivalent to

$$\lambda y_0 - Ay_0 - \mathcal{K}y_1 = z_0 \in D_A\left(\frac{1}{2}; 2\right),$$

$$\lambda y_1 - \frac{dy_1}{d\theta} = z_1 \in L^2(-\infty, 0; D(A)).$$

The second equation yields

$$y_1(\theta) = e^{\lambda\theta} y_1(0) + \int_{\theta}^0 e^{\lambda(\theta-\eta)} z_1(\eta) d\eta.$$

Setting $y_1(0) = y_0$ and substituting y_1 into the first equation, we obtain

$$(\lambda - \mathcal{A} - \mathcal{K}e^{\lambda\cdot})y_0 = z_0 + \mathcal{K} \int_{\theta}^0 e^{\lambda(\theta-\eta)} z_1(\eta) d\eta =: \bar{z}_0.$$

Noting that $\mathcal{K}e^{\lambda\cdot}y_0 = \tilde{K}(\lambda)y_0$, we have

$$(\lambda - \mathcal{A} - \tilde{K}(\lambda))y_0 = \bar{z}_0.$$

By virtue of Theorem 1.1(iv) this is solvable for any $\lambda > r_0$ and $y_0 = (\lambda - \mathcal{A} - \tilde{K}(\lambda))^{-1}\bar{z}_0 \in D(\mathcal{A})$. Then

$$y_1(\theta) = e^{\lambda\theta} (\lambda - \mathcal{A} - \tilde{K}(\lambda))^{-1}\bar{z}_0 + \int_{\theta}^0 e^{\lambda(\theta-\eta)} z_1(\eta) d\eta$$

lies in $W^{1,2}(-\infty, 0; D(A))$. We also have

$$Ay_0 + \mathcal{K}y_1 = \lambda y_0 - z_0 \in D_A\left(\frac{1}{2}; 2\right).$$

Hence $\lambda \in \rho(\mathcal{A})$ and \mathcal{A} is the infinitesimal generator of the semigroup $\underline{S}(t)$. \square

Remark 1.5. Let $A = A_0 + A_1$, where A_0 is selfadjoint and negative. If $A_1 \in L(D(-A)^{1/2}; H)$, then we can replace $D_A(\frac{1}{2}; 2)$ by $D(-A)^{1/2}$ and $D_A(-\frac{1}{2}; 2)$ by $D(-A^*)^{1/2}$ (see §2).

In [14] a special case of the delay equation in [15] was considered and a quadratic control problem on $D_A(\frac{1}{2}; 2) \times L^2(-r, 0; D(A))$ was solved.

If we take $B \in L(U; D_A(\frac{1}{2}; 2))$, $M \in L(D_A(\frac{1}{2}; 2), H_0)$, and $G \in L^+(D_A(\frac{1}{2}; 2))$, then by using the semigroup $\underline{S}(t)$ in Theorem 1.4 we can solve our control problem as in [14]; however, the state space \underline{Z} is not convenient in applications, and we wish to take the initial value y_0 in H rather than in $D_A(\frac{1}{2}; 2)$. Moreover, our cost functionals (2) or (3) are more natural, as we can see from examples (see Example 5.1). Thus we need a representation of our system (1) in a larger space.

2. The semigroup model. Let $D_A(-\alpha, 2)$, $\alpha \in]0, 1[$, be the extrapolation space of A (see [2]). To take y_0 in H rather than in $D_A(\frac{1}{2}; 2)$ we replace H (respectively, $D(A)$) by $D_A(-\frac{1}{2}; 2)$ (respectively, $D_A(\frac{1}{2}; 2)$) and assume, in addition to Hypothesis 1, the following hypothesis.

Hypothesis 3. $K(\cdot) \in L^2(0, \infty); L(D_A(-\frac{1}{2}; 2); D_A(\frac{1}{2}; 2))$.

Then the operator \mathcal{K} in (11) belongs to $L(L^2(-\infty, 0; D_A(\frac{1}{2}; 2)); D_A(-\frac{1}{2}; 2))$. By translation we obtain all results similar to those in §1. In particular, we state results corresponding to Corollary 1.3 and Theorem 1.4, respectively.

THEOREM 2.1. (i) *For each $y_0 \in H$ and $y_1 \in L^2(-\infty, 0; D_A(\frac{1}{2}; 2))$ there exists a unique solution to (10) in*

$$L^2(0, T; D_A(\frac{1}{2}; 2)) \cap W^{1,2}(0, T; D_A(-\frac{1}{2}; 2)) \subset C([0, T]; H).$$

(ii) *Assume Hypothesis 3, and let $y_1 \in W^{1,2}(-\infty, 0; D_A(\frac{1}{2}; 2))$, $y_1(0) = y_0$, and $Ay_0 + Ky_1 \in H$. Then there exists a unique solution to (9) in*

$$W^{1,2}(-\infty, T; D_A(\frac{1}{2}; 2)) \cap W^{2,2}(0, T; D_A(-\frac{1}{2}; 2)) \subset C^1([0, T]; H).$$

THEOREM 2.2. *Assume Hypothesis 1 and 3 with H (respectively $D(A)$) replaced by $D_A(-\frac{1}{2}; 2)$ (respectively $D_A(\frac{1}{2}; 2)$). Let $y_0 \in H$, let $y_1 \in L^2(-\infty, 0; D_A(\frac{1}{2}; 2))$, and let $y(t)$ be the solution of (9) (and hence of (13)) given in Theorem 2.1(i). Define the map on $Z = H \times L^2(-\infty, 0; D_A(\frac{1}{2}; 2))$*

$$(17) \quad S(t) : \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \rightarrow \begin{pmatrix} y(t) \\ y_t(\cdot) \end{pmatrix}.$$

Then $S(t)$ is a strongly continuous semigroup on Z , and its infinitesimal generator is given by

$$(18) \quad \mathcal{A} \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} = \begin{pmatrix} Ay_0 + Ky_1 \\ \frac{dy_1}{d\theta} \end{pmatrix}$$

$$(19) \quad D(\mathcal{A}) = \{(y_0, y_1) \in Z : y_1 \in W^{1,2}(-\infty, 0; D_A(\frac{1}{2}; 2)), y_1(0) = y_0, Ay_0 + Ky_1 \in H\}.$$

Next we express $S(t)$ by using the resolvent operator. We write (9) as

$$(20) \quad y'(t) = Ay(t) + \int_0^t K(t-r)y(r) dr + K_1(t)y_1,$$

where

$$K_1(t)y_1 = \int_{-\infty}^0 K(t-\theta)y_1(\theta) d\theta \in L^2\left(0, \infty; D_A\left(-\frac{1}{2}; 2\right)\right) \cap C\left([0, T]; D_A\left(-\frac{1}{2}; 2\right)\right).$$

The solution of (20) can be written as

$$y(t) = R(t)y_0 + \int_0^t R(t-r)K_1(r)y_1 dr$$

Set

$$S(t) = \begin{pmatrix} S_{11}(t) & S_{12}(t) \\ S_{21}(t) & S_{22}(t) \end{pmatrix};$$

then $y(t) = S_{11}(t)y_0 + S_{12}(t)y_1$. Thus we have

$$S_{11}(t)y_0 = R(t)y_0,$$

$$S_{12}(t)y_1 = \int_0^t R(t-r)K_1(r)y_1 dr.$$

Similarly, we have

$$(S_{21}(t)y_0)(\cdot) = R(t+\cdot)y_0,$$

$$(S_{22}(t)y_1)(\cdot) = \int_0^{t+\cdot} R(t-r)K_1(r)y_1 dr.$$

Hypotheses 1 and 3 come from physical examples such as Example 5.1. If we assume, instead of Hypothesis 3, the following hypothesis, we have, in fact, Corollary 2.3.

Hypothesis 3'. $K(\cdot) \in L^2(0, \infty; L(D_A(\frac{1}{2}; 2); H)) \cap L^2(0, \infty; L(H; D_A(-\frac{1}{2}; 2)))$. then we can find a semigroup on $H \times L^2(-\infty, 0; H)$.

COROLLARY 2.3. *Assume Hypothesis 3'. Then for each $y_0 \in H$ and $y_1 \in L^2(-\infty, 0; H)$ there exists a unique solution $y(t)$ to (13) in*

$$L^2(0, T; D_A(\frac{1}{2}; 2)) \cap C([0, T]; H)$$

for any $T > 0$. The map

$$(21) \quad S(t) : \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \rightarrow \begin{pmatrix} y(t) \\ y_t(\cdot) \end{pmatrix}$$

is a strongly continuous semigroup on $Z = H \times L^2(-\infty, 0; H)$. Its infinitesimal generator is given by

$$(22) \quad \mathcal{A} \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} = \begin{pmatrix} Ay_0 + Ky_1 \\ \frac{dy_1}{d\theta} \end{pmatrix},$$

$$(23) \quad D(\mathcal{A}) = \{(y_0, y_1) \in Z : y_1 \in W^{1,2}(-\infty, 0; H), y_1(0) = y_0, Ay_0 + Ky_1 \in H\}.$$

If $K(\cdot) \in L^2(0, \infty; H)$, then A need not be analytic.

Hypothesis 3''. $K(\cdot) \in L^2(0, \infty; L(H))$.

COROLLARY 2.4. *Let A be any infinitesimal generator of a strongly continuous semi-group on H . Assume Hypothesis 3''. Then for each $y_0 \in H$ and $y_1 \in L^2(-\infty, 0; H)$ there exists a unique solution $y(t)$ to (13) in $C([0, T]; H)$ for any $T > 0$. Define the map $S(t)$ as in (21). Then it is a strongly continuous semigroup on $Z = H \times L^2(-\infty, 0; H)$ with generator (22), (23).*

See [12] for more general cases of integrodifferential operators where A is not analytic.

3. Quadratic control on finite horizon. Now we consider (13) with control

$$(24) \quad \begin{cases} y'(t) = Ay(t) + Ky_t + Bu(t) \\ y(0) = y_0 \\ y(\theta) = y_1(\theta), \quad \theta \in]-\infty, 0], \end{cases}$$

where $y_1 \in L^2(-\infty, 0; D_A(\frac{1}{2}, 2))$. Then by setting $z(t) = [y(t), y_t(\cdot)]'$ we obtain

$$(25) \quad \begin{cases} z'(t) = Az(t) + \tilde{B}u(t), \\ z(0) = [y_0, y_1]', \end{cases}$$

where

$$\tilde{B} = \begin{pmatrix} B \\ 0 \end{pmatrix}.$$

For each $u \in L^2(0, T; U)$ we define the mild solution of (25) by

$$(26) \quad z(t) = S(t)[y_0, y_1]' + \int_0^t S(t-r)\tilde{B}u(r) dr.$$

The mild solution (6) of (1) corresponds to the first component of $z(t)$ of the special case $y_1 = 0$, i.e.,

$$(27) \quad z(t) = S(t)[y_0, 0] + \int_0^t S(t-r)\tilde{B}u(r) dr.$$

The cost functional (2) can be rewritten

$$(28) \quad J(u) = \int_0^T \left[|\tilde{M}z(t)|^2 + |u(t)|^2 \right] dt + \langle \tilde{G}z(T), z(T) \rangle,$$

where

$$\tilde{M} = \begin{pmatrix} M \\ 0 \end{pmatrix} \in L(Z; H_0) \quad \text{and} \quad \tilde{G} = \begin{pmatrix} G & 0 \\ 0 & 0 \end{pmatrix} \in L^+(Z).$$

The control problem (26), (28) is a standard quadratic problem [1] in the state-space form [8], [9]. As is well known, the optimal control is given by the feedback law

$$(29) \quad \underline{u} = -\tilde{B}^*Q(t)z(t),$$

where Q is the unique selfadjoint nonnegative solution of the Riccati equation

$$(30) \quad \begin{cases} Q' + \mathcal{A}^*Q + Q\mathcal{A} + \tilde{M}^*\tilde{M} - Q\tilde{B}\tilde{B}^*Q = 0, \\ Q(T) = \tilde{G}. \end{cases}$$

Setting

$$Q(t) = \begin{pmatrix} Q_{11}(t) & Q_{12}(t) \\ Q_{21}(t) & Q_{22}(t) \end{pmatrix},$$

we can write (29) in the form

$$(31) \quad \underline{u}(t) = -B^*[Q_{11}(t)y(t) + Q_{12}(t)y_t].$$

The minimal cost corresponding to \underline{u} is

$$(32) \quad J(\underline{u}) = \left\langle Q(0) \begin{pmatrix} y_0 \\ y_1 \end{pmatrix}, \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \right\rangle.$$

Hence the minimal cost for the problem (1), (2) is given by

$$(33) \quad J(\underline{u}) = \langle Q_{11}(0)y_0, y_0 \rangle.$$

Summing up, we have the following theorem.

THEOREM 3.1. *Assume Hypothesis 1 and 3 are true. Then there exists a unique optimal control for the problem (1), (2). It is given by the feedback law (31), and the minimal cost is given by (33).*

For the control problem (1), (2), where $K(\cdot)$ satisfies either Hypothesis 3' or Hypothesis 3'', the feedback law (31) is still optimal and the optimal cost is given by (33).

4. Quadratic control on infinite horizon. Here we consider the control problem (1), (3). To avoid the trivial case we make the following assumption for (27).

Hypothesis 4. For each $y_0 \in H$ and $y_1 \in L^2(-\infty, 0; D_A(\frac{1}{2}, 2))$ there exists a control $u \in L^2(0, \infty; U)$ such that

$$(34) \quad J(u) = \int_0^\infty [|\tilde{M}z(t)|^2 + |u(t)|^2] dt < \infty.$$

Later we give sufficient conditions for Hypothesis 4. Let $Q_T(t)$ be the solution of the Riccati equation (30) with $Q_T(T) = 0$. Then the following is known.

PROPOSITION 4.1. *Assume Hypotheses 3 and 4 are true. Then there exists a strong limit Q_∞ of Q_T . Q_∞ is the minimal nonnegative solution of the algebraic Riccati equation*

$$(35) \quad \mathcal{A}^*Q + Q\mathcal{A} + \tilde{M}^*\tilde{M} - Q\tilde{B}\tilde{B}^*Q = 0.$$

If \mathcal{A} , \widetilde{M} is detectable, then Q_∞ is the unique nonnegative solution of (35). Moreover, $A - \widetilde{B}\widetilde{B}^*Q_\infty$ generates an exponentially stable semigroup on Z .

THEOREM 4.2. Assume Hypotheses 3 and 4 are true. Then there exists a unique optimal control for (26), (34). It is given by the feedback law

$$(36) \quad \underline{u} = -\widetilde{B}^*Q_\infty z(t),$$

and the minimal cost is

$$(37) \quad J(\underline{u}) = \left\langle Q_\infty \begin{pmatrix} y_0 \\ y_1 \end{pmatrix}, \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \right\rangle.$$

In particular, the optimal control for the problem (1), (2) is given by

$$(38) \quad \underline{u}(t) = -B^*[Q_{11}y(t) + Q_{12}y_t]$$

and

$$(39) \quad J(\underline{u}) = \langle Q_{11}y_0, y_0 \rangle,$$

where

$$Q_\infty = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}.$$

If $Q_{12}(y_1) = \int_{-\infty}^0 Q_{12}(-\theta)y_1(\theta) d\theta$ for some $Q_{12} \in L^2(0, \infty; L(D_A(\frac{1}{2}, 2); H))$, then the optimal closed-loop system corresponding to (38) is

$$(40) \quad y'(t) = (A - BB^*Q_{11})y(t) + \int_0^t [K(t-r) - BB^*Q_{12}(t-r)]y(r) dr$$

and of the same form as (1). If $(\mathcal{A}, \widetilde{M})$ is detectable, then the resolvent operator for (40) is exponentially stable.

A sufficient condition for Hypothesis 4 can be found in [7]. For the sake of completeness we quote some results from [7]. Let $\widehat{K}(\cdot)$ be a maximal analytic extension of the Laplace transform of K , and let Ω_0 be its domain of definition. We set

$$\begin{cases} \rho_0 = \{\lambda \in \Omega : \exists (\lambda - A - \widehat{K}(\lambda))^{-1}\}, \\ F(\lambda) = (\lambda - A - \widehat{K}(\lambda))^{-1} \quad \text{for } \lambda \in \rho_0, \end{cases}$$

and we denote by ρ_1 the set of all isolated removable singularity of $F(\cdot)$. Moreover, we set

$$\begin{cases} \rho = \rho_0 \cup \rho_1, \\ F(\lambda) = \lim_{z \rightarrow \lambda} F(z), \quad \lambda \in \rho \setminus \rho_0. \end{cases}$$

Define the *generalized spectrum* $\sigma = C \setminus \rho$. If λ_0 is a pole of $F(\cdot)$ of order m_0 , we set, for λ sufficiently close to λ_0 ,

$$F(\lambda) = \sum_{n=0}^{\infty} S_n(\lambda - \lambda_0)^n + \sum_{n=0}^{m_0-1} Q_n(\lambda - \lambda_0)^{-n-1},$$

where

$$\begin{cases} Q_n = \frac{1}{2\pi i} \int_{C(\lambda_0, \varepsilon)} F(\lambda)(\lambda - \lambda_0)^n d\lambda, \\ S_n = \frac{1}{2\pi i} \int_{C(\lambda_0, \varepsilon)} F(\lambda)(\lambda - \lambda_0)^{-n-1} d\lambda, \end{cases}$$

and $C(\lambda_0, \varepsilon)$ is the circle with center λ_0 having sufficiently small radius $\varepsilon > 0$.

Let $\omega > 0$ be such that $\sigma \cap \{\lambda \in C : \operatorname{Re} \lambda = -\omega\} = \emptyset$, and let $\sigma_+(\omega) = \sigma \cap \{\lambda \in C : \operatorname{Re} \lambda > -\omega\}$, $\sigma_-(\omega) = \sigma \cap \{\lambda \in C : \operatorname{Re} \lambda < -\omega\}$. We can make the following assumption.

Hypothesis 5. (i) $\sigma_+(\omega) = \{\lambda_1, \dots, \lambda_N\}$, where for each $j = 1, \dots, N$, λ_j is a pole of $F(\cdot)$ of order $m_j < \infty$.

(ii) The residues $R_{j,k}$, $k = 0, 1, \dots, m_j$, of $F(\cdot)$ at $\lambda = \lambda_j$ are finite-rank operators.

The condition below is called a *Hautus condition*.

Hypothesis 6. Range $Q_{j,k}^* \cap \operatorname{Ker} B^* = \{0\}$ for all $j = 1, 2, \dots, N$ and $k = 0, 1, \dots, m_j$.

Let X be a Banach space, and let $C_\omega([0, \infty[; X)$ be the space of bounded continuous functions $x(t)$ in X with property $\sup_{t>0} \|x(t)e^{\omega t}\|_X < +\infty$. Under Hypothesis 5 it is shown [7] that Hypothesis 6 holds if and only if the following is true:

For each $y_0 \in H$ there exists a control $u \in C_\omega([0, \infty[; U)$ (in fact, $u \in C_\omega^\alpha([0, \infty[; U)$) such that $y \in C_\omega([0, \infty[; H)$, where y is the solution of (1). Hence if Hypotheses 5 and 6 hold, then the control problem (1), (3) is well defined. Modifying slightly the proof of [7, Thm. 2.3], we can show that under Hypothesis 6 system (13) is stabilizable in the above sense. We have, in fact, the following result, the proof of which was suggested to us by A. Lunardi.

THEOREM 4.3. *If Hypothesis 6 holds, then the system (13) is stabilizable, i.e., for each $y_0 \in H$ there exists a control $u \in C_\omega([0, \infty[; U)$ such that $y \in C_\omega([0, \infty[; H)$ for some $\omega \in]0, \omega_0[$.*

Proof. Define

$$R_{\lambda_j}(t) = \frac{1}{2\pi i} \int_{C(\lambda_j, \varepsilon)} e^{\lambda t} F(\lambda) d\lambda = \sum_{k=0}^{m_j-1} \frac{e^{\lambda_j t} t^k}{k!} Q_{j,k}$$

and

$$R_+^{\omega_0}(t) = \sum_{j=1}^N R_{\lambda_j}(t).$$

Then as in [7, Prop. 1.1] we can show that problem (13) is stabilizable in the above sense if and only if for each $y_0 \in H$ and $y_1 \in L^2(-\infty, 0, D_A(\frac{1}{2}, 2))$ there exists $u \in C_\omega([0, \infty[; U)$ such that

$$(41) R_+^\omega(t) y_0 + \int_0^{+\infty} R_+^\omega(t-s) K_1(s) y_1 ds = - \int_0^{+\infty} R_+^\omega(t-s) B u(s) ds, \quad t \geq 0,$$

where $K_1(\cdot)$ is as given in §2. First, we assume $\operatorname{Re} \lambda_j > 0$, $j = 1, 2, \dots, N$, and show (41) with $\omega = \omega_0$. If there exists λ_j with $\operatorname{Re} \lambda_j = 0$, then we can set $\underline{v}(t) = e^{\varepsilon t} y(t)$ for sufficiently small $\varepsilon > 0$ and reduce the problem to the case for which $\omega = \omega_0 - \varepsilon$.

As in the proof of Theorem 2.3 in [7], we can show that (41) is equivalent to

$$\begin{aligned}
 (42) \quad & \sum_{j=1}^N \sum_{n=0}^{m_j-1} e^{\lambda_j t} \frac{t^n}{n!} Q_{j,n} y_0 + \sum_{j=1}^N \sum_{n=0}^{m_j-1} \sum_{k=n}^{m_j-1} e^{\lambda_j t} \frac{t^n}{n!} \int_0^{+\infty} e^{-\lambda_j s} (-s)^{k-n} Q_{j,k} K_1(s) y_1 ds \\
 &= - \sum_{j=1}^N \sum_{n=0}^{m_j-1} \sum_{k=n}^{m_j-1} e^{\lambda_j t} \frac{t^n}{n!} \int_0^{+\infty} e^{-\lambda_j s} (-s)^{k-n} Q_{j,k} B u(s) ds.
 \end{aligned}$$

Note that the second term is well defined since $K_1(\cdot)y_1$ is bounded and $\operatorname{Re} \lambda_j > 0$. Since the functions $t \rightarrow e^{\lambda_j t} t^n$ are linearly independent, (42) is equivalent to $\Gamma u = Q[y_0, K_1(\cdot)y_1]$, where

$$\Gamma : C_\omega([0, +\infty[; U) \rightarrow H^K, \quad K = \sum_{j=1}^N m_j,$$

$$\Gamma u = \left\{ \sum_{k=n}^{m_j-1} \int_0^{+\infty} e^{-\lambda_j s} (-s)^{k-n} Q_{j,k} B u(s) ds \right\}_{j=1, \dots, N; n=0, \dots, m_j-1},$$

and $Q : H \rightarrow H^K$,

$$\begin{aligned}
 & Q(y_0, K_1(\cdot)y_1) \\
 &= \left\{ Q_{j,n} y + \sum_{k=n}^{m_j-1} \int_0^{+\infty} e^{-\lambda_j s} (-s)^{k-n} Q_{j,k} K_1(s) y_1 ds \right\}_{j=1, \dots, N; n=0, \dots, m_j-1}.
 \end{aligned}$$

Since the range of Γ and Q are finite-dimensional, (42) holds if and only if

$$(43) \quad \operatorname{Ker} Q^* \supset \operatorname{Ker} \Gamma^*.$$

For each $(h_{jn}) = (h_{jn})_{j=1, \dots, N; n=0, \dots, m_j-1} \in H^K$ we have

$$\begin{aligned}
 \Gamma^*(h_{jn})u &= - \sum_{j=1}^N \sum_{k=0}^{m_j-1} \sum_{h=0}^{m_j-1-k} \int_0^{+\infty} \frac{(-s)^k}{k!} e^{-\lambda_j s} \langle u(s), B^* Q_{j,k+h}^* h_{j,k+h} \rangle ds \\
 &\quad + Q^*(h_{jn})(y_0, K_1(\cdot)y_1) \\
 &= \sum_{j=1}^N \sum_{h=0}^{m_j-1} \langle y_0, Q_{jh}^* h_{jh} \rangle \\
 &\quad + \sum_{j=1}^N \sum_{k=0}^{m_j-1} \sum_{h=0}^{m_j-1-k} \int_0^{+\infty} \frac{(-s)^k}{k!} e^{-\lambda_j s} \langle K_1(s) y_1, Q_{j,k+n}^* h_{j,k+n} \rangle ds
 \end{aligned}$$

so that

$$\begin{aligned}
 \operatorname{Ker} \Gamma^* &= \left\{ (h_{jn}) \in H^K : B^* \left(\sum_{h=k}^{m_j-1} Q_{jh}^* h_{jh} \right), j = 1, \dots, N, k = 0, 1, \dots, m_j - 1 \right\} \\
 &= \{ (h_{jn}) \in H^K : B^* Q_{jh}^* h_{jn}, j = 1, \dots, N, k = 0, 1, \dots, m_j - 1 \},
 \end{aligned}$$

$$\text{Ker } Q^* \supset \{(h_{jn}) \in H^K : \sum_{j=1}^N \sum_{h=0}^{m_j-1} Q_{jh}^* h_{jn} = 0\}.$$

Now we can show that (13) is stabilizable and is equivalent to (43). It is easy to see that if (13) is stabilizable, then (43) holds. Conversely, assume that (23) holds and let $h \in \text{Ker } B^* Q_{j_0 h_0}^*$ for some j_0, h_0 . Set $h_{jn} = \delta_{j,j_0} \delta_{n,n_0} h$; then $B^* Q_{jn}^* h_{jn} = 0$ for each $j = 1, \dots, N, h = 0, \dots, m_j - 1$. By (43) we have

$$\sum_{j=1}^N \sum_{h=0}^{m_j-1} Q_{jh}^* h_{jn} = Q_{j_0 n_0}^* h = 0.$$

Therefore, for each j_0, h_0 we have

$$\text{Ker } Q_{j_0 n_0}^* \supset B^* Q_{j_0 n_0}^*,$$

and (13) is stabilizable. \square

Now consider the detectability of $\mathcal{A}, \widetilde{M}$. It is useful to consider the following:

$$(44) \quad \begin{cases} \underline{\eta}'(t) = -A^* \underline{\eta}(t) - \int_t^T K^*(r-t) \underline{\eta}(r) dr, \\ \underline{\eta}(T) = \eta_1. \end{cases}$$

Suppose we have classical solutions for (4) and (44). Then by differentiating $\langle y(t), \underline{\eta}(t) \rangle$, integrating from $t = 0$ to T , and using the Fubini theorem we obtain

$$\langle y(T), \eta_1 \rangle = \langle y_0, \underline{\eta}(0) \rangle.$$

Hence (44) is the adjoint system of (4). Equations (44) can be also written

$$(45) \quad \begin{cases} \eta'(t) = A^* \eta(t) + \int_0^t K^*(t-r) \eta(r) dr, \\ \eta(0) = \eta_1. \end{cases}$$

Thus the detectability of $(\mathcal{A}, \widetilde{M})$ is translated into the stabilizability of the following system:

$$(46) \quad \eta'(t) = A^* \eta(t) + \int_0^t K^*(t-r) \eta(r) dr + \widetilde{M} v(t).$$

If A^* and K^* have properties similar to those of A and K we can obtain sufficient conditions for detectability. Finally, we note that we can also solve control problems for inhomogeneous systems as in [4], and [5].

Example 4.4. Let Ω be a bounded open domain in R^n with C^2 boundary $\partial\Omega$. Consider the heat equation in materials of the *fading-memory* type introduced by Nunziato [20]:

$$(47) \quad \begin{cases} b_0 \frac{\partial y}{\partial t}(t, x) + \frac{\partial}{\partial t} \int_0^t \beta(t-r) y(r, x) dr \\ \quad = c_0 \Delta y(t, x) - \int_0^t \gamma(t-r) \Delta y(r, x) dr + B_0 u(t, x), \quad t > 0, x \in \overline{\Omega}, \\ y(0, x) = y_0(x), \quad x \in \overline{\Omega}, \\ \Gamma y(t, x) = 0, \quad t > 0, \quad x \in \partial\Omega, \end{cases}$$

where $\Gamma y = y$ or $\Gamma y = \partial y / \partial n$. $y(t, x)$ represents the temperature at $x \in \bar{\Omega}$ at time t , b_0 , and c_0 are positive constants, and u is the heat supply. β and γ are completely monotone kernels with

$$\beta(t) = \int_0^\infty e^{-\omega t} \mu(d\omega), \quad \gamma(t) = \int_0^\infty e^{-\omega t} \nu(d\omega),$$

where μ and ν are positive Borel measures with compact support $\text{supp } \mu$ and $\text{supp } \nu$ contained in $]a, \infty[$, $a > 0$. Then the heat equation (47) can be written as (1) in $H = L^2(\Omega)$ with

$$Ah = \frac{1}{b_0}(c_0 \Delta h - \beta(0)h), \quad D(A) = \{h \in H : \Delta h \in H, \Gamma h = 0\},$$

$$K(t)h = \frac{1}{b_0}(-\beta'(t)h - \gamma(t)\Delta h),$$

$$B = \frac{1}{b_0}B_0.$$

Let $\tilde{\beta}(\cdot)$ and $\tilde{\gamma}(\cdot)$ be analytic extensions of the Laplace transforms of β and γ to $C \setminus \text{supp } \mu$ and $C \setminus \text{supp } \nu$, respectively. Then

$$\rho_0 = \left\{ \lambda \in C : c_0 - \tilde{\gamma}(\lambda) \neq 0, \frac{\lambda(b_0 + \tilde{\beta}(\lambda))}{c_0 - \tilde{\gamma}(\lambda)} \neq -\lambda_n \right\},$$

$$F(\lambda) = \frac{b_0}{\frac{\lambda(b_0 + \tilde{\beta}(\lambda))}{c_0 - \tilde{\gamma}(\lambda)}} R \left(\frac{\lambda(b_0 + \tilde{\beta}(\lambda))}{c_0 - \tilde{\gamma}(\lambda)}, \Delta \right),$$

where $\{-\lambda_n\}$ is the decreasing sequence of the eigenvalues of Δ . Hence

$$\{\lambda \in: \text{Re } \lambda \geq 0\} = \begin{cases} \emptyset & \text{if } \Gamma y = y \text{ on } \partial\Omega, \\ \{0\} & \text{if } \Gamma y = \frac{\partial y}{\partial n} \text{ on } \partial\Omega. \end{cases}$$

If $\Gamma y = y$, then (47) with $u = 0$ is stable. Hence our control problems (1), (2) and (1), (3) are well defined. If $\Gamma y = \frac{\partial y}{\partial n}$, then (47) with $u = 0$ is not stable, but satisfies Hypothesis 5. This implies that if $h(x) = h_0$ (constant different from 0) $\notin \text{Ker } B^*$, then (47) is stabilizable. If $Bu = b(x)u(t)$, $b(\cdot) \in H$, and u is a scalar, then $\int_\Omega b(x) dx \neq 0$ implies stabilizability. Hence the quadratic problems (1), (3) and, of course, (1), (2) are well defined. See [6] for more discussions on this example and other examples of system (1). See also [19] for heat equations with memory.

The following example was introduced to us by J. Zabczyk and is covered by our model.

Example 4.5. Consider the delay equation

$$(48) \quad \begin{cases} x''(t) = -k(0)x(t) - \int_{-\infty}^0 k'(-r)x(t+r) dr + u(t), \\ x(0) = x_0, \quad x'(0) = x_1, \end{cases}$$

which can be regarded as a model of the oscillation of a particle suspended by light linearly viscoelastic string, where the relaxation modulus $k : [0, \infty[\rightarrow \mathbb{R}$ is differentiable and such that

$$k(t) > 0, \quad k'(t) < 0, \quad k''(t) > 0, \quad \lim_{t \rightarrow \infty} k'(t) = 0.$$

Setting

$$y = \begin{pmatrix} x \\ x' \end{pmatrix},$$

we can write (48) as

$$y'(t) = \begin{pmatrix} 0 & 1 \\ -k(0) & 0 \end{pmatrix} y(t) - \int_0^t k(t-r) \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} y(r) dr + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t),$$

$$y(0) = y_0 = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix}.$$

If $k \in L^2(0, \infty)$, then this is a special case covered by Theorem 4.2. If $k(t) = e^{-at}$, $a > 0$, the spectrum σ is given by

$$\sigma = \{\lambda : \lambda^3 + a\lambda^2 + k(0)\lambda + ak(0) - 1 = 0\}.$$

Set

$$A = \begin{pmatrix} 0 & 1 \\ -k(0) & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad M = [1, 0];$$

then (A, B) is controllable and (M, A) is observable. Hence the minimization problem

$$J(u) = \int_0^\infty [|x(t)|^2 + |u(t)|^2] dt$$

is well defined.

REFERENCES

- [1] R. F. CURTAIN AND A. J. PRITCHARD (1978), *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin.
- [2] G. DA PRATO AND P. GRISVARD (1984), *Maximal regularity for evolution equations by interpolation and extrapolation*, J. Funct. Anal., 58, pp. 107-124.
- [3] G. DA PRATO AND M. IANNELLI (1985), *Existence and regularity for a class of integrodifferential equations of parabolic type*, J. Math. Anal. Appl., 112, pp. 35-55.
- [4] G. DA PRATO AND A. ICHIKAWA (1988), *Optimal control for linear periodic systems*, Appl. Math. Optim., 18, pp. 39-66.
- [5] ——— (1990), *Quadratic control for linear time varying systems*, SIAM J. Control Optim., 28, pp. 359-381.
- [6] G. DA PRATO AND A. LUNARDI (1988), *Solvability on the real line of a class of linear Volterra integrodifferential equations of parabolic type*, Ann. Mat. Pura Appl. (4), 150, pp. 67-118.
- [7] ——— (1990), *Stabilizability of integrodifferential parabolic equations*, J. Integral Equations Appl., 2, pp. 281-304.
- [8] M. C. DELFOUR (1986), *The linear-quadratic optimal control problem with delays in state and control variables: a state space approach*, SIAM J. Control Optim., 24, pp. 835-883.

- [9] M. C. DELFOUR AND J. KARRAKCHOU (1987), *State space theory of linear time invariant systems with delays in state, control and observation variables*, J. Math. Anal. Appl., 125, pp. 361–450.
- [10] M. C. DELFOUR, C. MCCALLA, AND S. K. MITTER (1975), *Stability and infinite-time quadratic cost problem for linear hereditary differential systems*, SIAM J. Control Optim., 13, pp. 48–88.
- [11] M. C. DELFOUR AND S. K. MITTER (1974), *Controllability, observability and optimal feedback control of hereditary differential systems*, SIAM J. Control Optim., 10, pp. 298–328.
- [12] W. DESCH, R. G. GRIMMER, AND W. SCHAPPACHER (1986), *Wellposedness and wave propagation for a class of integrodifferential equations*, Report 82, Institute für Mathematik, Universität Graz, Graz, Austria.
- [13] W. DESCH AND W. SCHAPPACHER (1985), *A semigroup approach to integrodifferential equations in Banach spaces*, J. Integral Equations, 10, pp. 99–110.
- [14] G. DI BLASIO (1981), *The linear quadratic optimal control problem for delay differential equations*, Rend. Acc. Naz. Lincei, 71, pp. 156–161.
- [15] G. DI BLASIO, KUNISCH, AND E. SINISTRARI (1984), *L^2 -regularity for parabolic partial integrodifferential equations with delay in the highest order derivative*, J. Math. Anal. Appl., 102, pp. 38–57.
- [16] A. LUNARDI (1985), *Laplace transform methods in integrodifferential equations*, J. Integral Equations, 10, pp. 185–211.
- [17] J. L. LIONS AND E. MAGENES (1968), *Problemes aux limites non homogenes et applications*, Dunod, Paris.
- [18] R. K. MILLER (1974), *Linear Volterra integrodifferential equations as semigroups*, Funkcial. Ekvac., 17, pp. 39–55.
- [19] ——— (1979), *An integrodifferential equation for rigid heat conductors with memory*, J. Math. Anal. Appl., 66, pp. 313–332.
- [20] J. W. NUNZIATO (1971), *On heat conduction in materials with memory*, Quart. Appl. Math., 29, pp. 187–304.

OPTIMAL CONTROL OF SWITCHING DIFFUSIONS WITH APPLICATION TO FLEXIBLE MANUFACTURING SYSTEMS*

MRINAL K. GHOSH[†], ARISTOTLE ARAPOSTATHIS[‡], AND STEVEN I. MARCUS[§]

Abstract. A controlled switching diffusion model is developed to study the hierarchical control of flexible manufacturing systems. The existence of a homogeneous Markov nonrandomized optimal policy is established by a convex analytic method. Using the existence of such a policy, the existence of a unique solution in a certain class to the associated Hamilton–Jacobi–Bellman equations is established and the optimal policy is characterized as a minimizing selector of an appropriate Hamiltonian.

Key words. flexible manufacturing system, Wiener process, switching diffusion, Poisson measure, Markov policy, dynamic programming equations

AMS subject classification. 93E20

1. Introduction. We study a controlled switching diffusion process that arises in numerous applications of systems with multiple modes or failure modes, including the hierarchical control of flexible manufacturing systems. A flexible manufacturing system (FMS) consists of a set of workstations capable of performing a number of different operations and interconnected by a transportation mechanism. An FMS produces a family of parts related by similar operational requirements or by belonging to the same final assembly [27]. The rapidly growing range of applicability of FMS includes metal cutting, assembly of printed circuit boards, integrated circuit fabrication, automobile assembly lines, etc. Due to their tremendous flexibility, FMSs are significantly more efficient in many ways than traditional manufacturing systems. However, the high capital cost of an FMS demands very efficient management of production and maintenance (repair/replacement) scheduling so that uncertain events can be taken care of such as random demand fluctuations, machine failures, inventory spoilage, sales returns, etc. The large size of the system and its associated complexities make it imperative to divide the control or management into a hierarchy consisting of a number of levels. Thus, the overall complex problem is reduced to a number of manageable subproblems at each level, and these levels are linked by means of a hierarchical integrative system. We refer to [1], [21], and [27] for a detailed description of these hierarchical schemes. We will confine our attention to the top two levels.

(i) generation of decision tables, which is accomplished by developing a suitable mathematical model describing the dynamical evolution of the system. This is done off-line.

*Received by the editors January 14, 1991; accepted for publication (in revised form) March 12, 1992. This research was supported in part by Texas Advanced Research Program (Advanced Technology Program) grants 003658-093 and 003658-186, in part by Air Force Office of Scientific Research grants AFOSR-91-0033, F49620-92-J-0045, and F49620-92-J-0083, and in part by National Science Foundation grant CDR-8803012.

[†]Department of Mathematics, Indian Institute of Science, Bangalore, India 560012.

[‡]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712.

[§]Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742.

(ii) The flow control level which plays the central role in the system. It determines, on line, the production and maintenance scheduling and continuously feeds the routing control level that calculates route splits, and that, in turn, governs the sequence controller which determines the scheduling times at which to dispatch parts. Since the top two levels directly govern the rest, it is of paramount importance to develop and study an appropriate mathematical model that will facilitate finding on-line, implementable, optimal feedback policies.

We first present a heuristic description of our model, which is a modified version of the model in [1], [21], and [27]. The FMS consists of L workstations, with each workstation having a number L_m of identical machines ($m = 1, 2, \dots, L$). A family of N types of different parts is produced. Let $u(t) = [u_1(t), \dots, u_N(t)]^T \in \mathbb{R}^N$ and $d(t) = [d_1(t), \dots, d_N(t)]^T \in \mathbb{R}^N$ denote the production rate (a control variable) and the downstream demand rate vectors of this family of parts, respectively. Also, $X(t) = [X_1(t), \dots, X_N(t)]^T \in \mathbb{R}^N$ denotes the downstream buffer stock. A negative value of $X_j(t)$, $j = 1, \dots, N$, indicates a backlogged demand for part j , while a positive value is the size of the inventory stored in the buffers. The evolution of $X(t)$ is governed by stochastic differential equations

$$(1.1) \quad \frac{dX(t)}{dt} = u(t) - d(t) + \text{diag}(\sigma_1, \dots, \sigma_N)\xi(t),$$

where $\sigma_i > 0$, $i = 1, \dots, N$ and $\xi(t) = [\xi_1(t), \dots, \xi_N(t)]^T$ is an N -dimensional white noise which can be interpreted as sales returns, inventory spoilage, sudden demand fluctuations, etc. (see [8]).

If $S_m(t)$ denotes the number of operational machines in station m at time t , then the state of the workstations may be represented by the L -tuple

$$S(t) = (S_1(t), \dots, S_L(t)).$$

The evolution of $S(t)$ is influenced by the inventory size and production scheduling, and can also be controlled by various decisions such as to produce, repair, or replace. The dynamics of $S(t)$ can be described as follows:

$$(1.2) \quad P\{S_m(t + \delta t) = \ell + 1 \mid S_m(t) = \ell\} \\ = \begin{cases} (L_m - \ell)v_m(t)\delta t + o(\delta t) & \text{for } 0 \leq \ell < L_m, \\ 0 & \text{otherwise,} \end{cases}$$

where $v_m(t)$, $m = 1, \dots, L$, are suitable control variables. In the uncontrolled case, $v_m(t) = \gamma_m$, which represents the infinitesimal repair rate at station m . These repair rates may depend implicitly on $X(t)$. This model also allows for a control variable reflecting the decision to repair or to replace on the basis of the inventory size. Also,

$$(1.3) \quad P\{S_m(t + \delta t) = \ell - 1 \mid S_m(t) = \ell\} \\ = \begin{cases} \ell p_m(X(t), u(t))\delta t + o(\delta t) & \text{for } 0 \leq \ell < L_m, \\ 0 & \text{otherwise,} \end{cases}$$

where p_m models the infinitesimal failure rate at the m th station. Equations (1.2) and (1.3) imply that

$$P\{S_m(t + \delta t) = \ell_1 \mid S_m(t) = \ell_2\} = 0 \quad \text{for } |\ell_1 - \ell_2| > 1.$$

With i and j denoting two states of the system, we define

$$\lambda_{ij}(\cdot)\delta t + o(\delta t) = P\{S(t + \delta t) = j \mid S(t) = i\}, \quad i \neq j$$

and

$$\lambda_{ii}(\cdot) = - \sum_{j \neq i} \lambda_{ij}(\cdot).$$

The machine state $S(t)$ can thus be modeled as a continuous time-controlled jump process taking values in a finite state space. In the uncontrolled case, $S(t)$ becomes a continuous time homogeneous Markov chain with infinitesimal generator given by the matrix $[\lambda_{ij}]$.

The choice of the production rate at each instant is constrained by the capacity of the currently operational machines. This translates into the requirement that at each time t the production rates must lie in some set $\Gamma(S(t))$ that depends on the machine state.

Let $y_{mn}^k(t)$ be the number of type n parts that undergo operation k at the m th station per unit interval of time and $\tau_{mn}^k(t)$ the length of time required for the completion of this operation. The product $y_{mn}^k(t)\tau_{mn}^k(t)$ is the portion of each unit time interval that one or more operational machines at station m must dedicate to perform operation k on type n parts, as dictated by the flow rate $y_{mn}^k(t)$. Since the amount of work completed at each station per unit time interval cannot exceed the time available at the operational machines, the following constraint applies:

$$(1.4) \quad \sum_n \sum_k y_{mn}^k(t)\tau_{mn}^k(t) \leq S_m(t) \quad \text{for all } m.$$

Also, assuming that no material is allowed to accumulate within the system, the throughput $u_n(t)$ of type n parts must satisfy

$$(1.5) \quad u_n(t) = \sum_m y_{mn}^k(t) \quad \text{for all } k \text{ and } n.$$

Therefore, for each state i , the set $\Gamma(i)$ is defined as the collection of all production rates $u = [u_1, \dots, u_N]^T$ for which, with the machine state $S(t) = i$, there exist feasible flow rates $y_{mn}^k(t)$ satisfying (1.4) and (1.5).

The flow control problem can now be stated. Given an initial buffer state $X(0) = x$ and machine state $S(0) = i$, we wish to specify a production plan and maintenance (repair/replacement) policy that minimizes the performance index

$$(1.6) \quad J(x, i, u, v) = E \left[\int_0^\infty e^{-\alpha t} c(X(t), S(t), u(t), v(t)) dt \mid X(0) = x, S(0) = i \right],$$

where $c(\cdot)$ is a cost function, $\alpha > 0$ is a discount factor, $u(\cdot)$ is the production rate, and $v(\cdot)$ is the maintenance rate. The objective is to find $u(\cdot)$, $v(\cdot)$ for which the minimum is achieved in (1.6). The ideal production and maintenance policy for a wide class of cost functions would minimize J by producing parts at exactly the demand rate, thereby keeping the buffer at zero. Such a policy is generally impossible because of the failures of the machines and various other uncertainties.

This FMS model motivates the study of a stochastic optimization problem in a more abstract setting that subsumes the flow control problem in the FMS as a special

case. This abstract problem is manifested in numerous other situations. In [17] it is encountered in a hybrid model proposed for the study of dynamic phenomena in large scale interconnected power networks. Sworder [39], [40] describes possible applications to macroeconomic models and dynamic renewal problems in general. In addition, it should be useful at other levels of the hierarchy described in [21].

We will briefly describe this problem formally; a rigorous description will be given in §2. Let $\mathcal{S} = \{1, 2, \dots, M\}$ and let U_i , $i = 1, \dots, M$, be prescribed compact metric spaces. For each $i, j \in \mathcal{S}$, let $b(\cdot, \cdot, i, \cdot) : \mathbb{R}_+ \times \mathbb{R}^N \times U_i \rightarrow \mathbb{R}^N$ and $\lambda_{ij} : \mathbb{R}_+ \times \mathbb{R}^N \times U_i \rightarrow \mathbb{R}$, satisfying $\lambda_{ij} \geq 0$, for $i \neq j$ and $\sum_j \lambda_{ij}(\cdot) = 0$. A stochastic process $(X(t), S(t))$ taking values in $\mathbb{R}^N \times \mathcal{S}$ is given by

$$(1.7) \quad X(t) = X(0) + \int_0^t b(\tau, X(\tau), S(\tau), u(\tau)) d\tau + \text{diag}(\sigma_1, \dots, \sigma_N)W(t),$$

$$(1.8) \quad P\{S(t + \delta t) = j \mid S(t) = i, X(s), S(s), s \leq t\} = \lambda_{ij}(t, X(t), u(t))\delta t + o(\delta t),$$

where $\sigma_i > 0$, $i = 1, \dots, N$, are constants and $W(\cdot) = [W_1(\cdot), \dots, W_N(\cdot)]^T$ is an N -dimensional standard Brownian motion. The control $u(\cdot)$ is a $U := \prod_{i=1}^N U_i$ -valued process such that when $S(t) = i$, $u(\cdot)$ takes values in U_i and $u(\cdot)$ is nonanticipative with respect to the driving Brownian motion $W(t)$. Let $c : \mathbb{R}_+ \times \mathbb{R}^N \times \mathcal{S} \times U \rightarrow \mathbb{R}_+$ be the cost function and $\alpha > 0$ a prescribed discount factor. Define a cost functional of the form

$$(1.9) \quad E \left[\int_0^\infty e^{-\alpha t} c(t, X(t), S(t), u(t)) dt \right].$$

The objective is to find an optimal control policy $u(\cdot)$ that minimizes (1.9) and takes the feedback form $u(t) = \bar{v}(t, X(t), S(t))$ for a suitably defined function \bar{v} . In the next section, we will assume appropriate conditions on b and λ which will guarantee that (1.7), (1.8) are well defined. We note here that for a performance index of the form (1.9), m, λ, c may be assumed to be independent of t without any loss of generality. Also, by replacing each U_k by $\prod_{k=1}^M U_k$ and $b(\cdot, i, \cdot)$ by its composition with the projection $\prod_{k=1}^M U_k \rightarrow U_i$, we may assume that each U_i is a replica of a fixed compact metric space.

We now briefly mention some earlier work leading to ours. The class of controlled piecewise deterministic models with jump Markov disturbances have been studied by Sworder [38], Rishel [35], Oldser and Suri [33], Davis [19], and Vermes [42] among many others. The piecewise deterministic FMS model has been studied by Kimenia and Gershwin [27], who have developed a heuristic numerical method based on the maximum principle established in [35]. Akella and Kumar [2] have studied a simplified model and obtained explicit solutions for one machine producing a single commodity. In all these papers the jump process is modeled as a continuous time (uncontrolled) Markov chain. Boukas and Haurie [14], [15] have modified the FMS model of Kimenia and Gershwin by introducing new state variables describing machine wear as well as a control parameter in the jump process; their model incorporates preventive maintenance. They have obtained a maximum principle, thereby extending Rishel's formalism in [35]. They have also considered piecewise deterministic models. To obtain an optimal policy of the feedback type in these models we must impose very strong conditions on terms like b, λ governing the system and stringent restrictions on the set of allowable policies. At the same time, it is assumed in these models that

between any two successive jumps of $S(t)$, the dynamics governing $X(t)$ are deterministic. Thus, certain unavoidable environmental uncertainties are not taken into account. These factors restrict the scope of applicability of these models. We have tried to circumvent these difficulties by adding an additive noise term in the state dynamics. This is specifically done to take into account the various sources of environmental randomness. Addition of this noise removes practically all restrictions imposed on the set of allowable control policies, thereby substantially enhancing the range of its applicability. The switching diffusion problem has also been studied by Bensoussan and Lions [7], using a martingale problem formulation. However, our motivation and approach are quite different. In [7], it is assumed that for some $\delta > 0$, $-\lambda_{ii} > \delta > 0$, for each i . We have, instead, used a strong formulation which is very important for practical applications. In our formulation we do not need the condition $-\lambda_{ii} > \delta > 0$. We also refer to [6], [8], [9], [16], [20], [32], [36], [37], and [43] for related work.

Our paper is structured as follows. A rigorous description of the mathematical model of the FMS is given in §2. The optimization problem is formulated and subsequently reduced to an equivalent convex optimization problem via the study of associated occupation measures. The compactness of laws is established in §3, the convexity and extremality of occupation measures are studied in §4, and the proof of existence of optimal policies is given in §5. Section 6 deals with the characterization of optimal policies via dynamic programming equations. In §7, we apply our theory to a simplified model and derive some interesting results. Finally, §8 contains some concluding remarks. Note that we have used a convex analytic approach for this problem, as opposed to the traditional analytic one. For the discounted cost criterion, the latter approach is more economical and is sketched in the Appendix. However, the convex analytic approach is interesting in its own right and would be more flexible and powerful for certain other purposes, e.g., the pathwise average cost problem or problems with several constraints where the analytic approach does not seem to be amenable. For (nonswitching) controlled diffusions, these problems have been treated in [11, Chap. VI] and [12] by a convex analytic approach. We hope our approach to switching diffusions would be useful in various other situations.

2. Mathematical model and preliminaries. Let U be a compact metric space and $\mathcal{S} = \{1, \dots, M\}$. Let $\bar{b} = [\bar{b}_1, \dots, \bar{b}_N]^T : \mathbb{R}^N \times \mathcal{S} \times U \rightarrow \mathbb{R}^N$. For each $i \in \mathcal{S}$, $\bar{b}(\cdot, i, \cdot)$ is assumed to be bounded, continuous, and Lipschitz in its first argument uniformly with respect to the third. For $i, j \in \mathcal{S}$, let $\bar{\lambda}_{ij} : \mathbb{R}^N \times U \rightarrow \mathbb{R}$ be bounded, continuous, and Lipschitz in its first argument uniformly with respect to the second. Also, assume that for $i, j \in \mathcal{S}$, $i \neq j$, $\bar{\lambda}_{ij} \geq 0$, and $\sum_{j=1}^M \bar{\lambda}_{ij} = 0$, for any $i \in \mathcal{S}$. Let $\sigma_i > 0$, $i = 1, 2, \dots, N$, be prescribed numbers. For a Polish space Y , $\mathfrak{B}(Y)$ will denote its Borel σ -field and $\mathcal{P}(Y)$ the space of probability measures on $\mathfrak{B}(Y)$ endowed with the Prohorov topology, i.e., the topology of weak convergence [10]. Let $\mathfrak{M}(Y)$ be the set of all nonnegative integer-valued, σ -finite measures on $\mathfrak{B}(Y)$. Let $\mathfrak{M}_\sigma(Y)$ be the smallest σ -field on $\mathfrak{M}(Y)$ with respect to which all maps from $\mathfrak{M}(Y)$ into $\mathbb{N} \cup \{\infty\}$ of the form $\mu \mapsto \mu(B)$, with $B \in \mathfrak{B}(Y)$, are measurable. $\mathfrak{M}(Y)$ will always be assumed to be endowed with this measurability structure. Let $\mathcal{V} = \mathcal{P}(U)$ and $b = [b_1, \dots, b_N]^T : \mathbb{R}^N \times \mathcal{S} \times \mathcal{V} \rightarrow \mathbb{R}^N$ be defined by

$$(2.1) \quad b_i(\cdot, \cdot, v) := \int_U \bar{b}_i(\cdot, \cdot, u) v(du), \quad v \in \mathcal{V}, \quad i = 1, \dots, N.$$

Similarly, for $i, j \in \mathcal{S}$, $\lambda_{ij} : \mathbb{R}^N \times \mathcal{V} \rightarrow \mathbb{R}$ is defined as

$$(2.2) \quad \lambda_{ij}(\cdot, v) := \int_U \bar{\lambda}_{ij}(\cdot, u) v(du), \quad v \in \mathcal{V}, \quad i, j \in \mathcal{S}.$$

For $i, j \in \mathcal{S}$, $x \in \mathbb{R}^N$, and $v \in \mathcal{V}$, we construct the intervals $\Delta_{ij}(x, v)$ of the real line in the following manner (see also [13], [17]):

$$\begin{aligned} \Delta_{12}(x, v) &= [0, \lambda_{12}(x, v)), \\ \Delta_{13}(x, v) &= [\lambda_{12}(x, v), \lambda_{12}(x, v) + \lambda_{13}(x, v)), \\ &\vdots \\ \Delta_{1M}(x, v) &= \left[\sum_{j=2}^{M-1} \lambda_{1j}(x, v), \sum_{j=2}^M \lambda_{1j}(x, v) \right), \\ \Delta_{21}(x, v) &= \left[\sum_{j=2}^M \lambda_{1j}(x, v), \sum_{j=2}^M \lambda_{1j}(x, v) + \lambda_{21}(x, v) \right), \\ &\vdots \\ \Delta_{2M}(x, v) &= \left[\sum_{j=2}^M \lambda_{1j}(x, v) + \sum_{\substack{j=1 \\ j \neq 2}}^{M-1} \lambda_{2j}(x, v), \sum_{j=2}^M \lambda_{1j}(x, v) + \sum_{\substack{j=1 \\ j \neq 2}}^M \lambda_{2j}(x, v) \right), \end{aligned}$$

and so on. For fixed x and v , these are disjoint intervals, and the length of $\Delta_{ij}(x, v)$ is $\lambda_{ij}(x, v)$. Now define a function $h : \mathbb{R}^N \times \mathcal{S} \times \mathcal{V} \times \mathbb{R} \rightarrow \mathbb{R}$ by

$$(2.3) \quad h(x, i, v, z) = \begin{cases} j - i & \text{if } z \in \Delta_{ij}(x, v), \\ 0 & \text{otherwise.} \end{cases}$$

Let $(X(t), S(t))$ be the $(\mathbb{R}^N \times \mathcal{S})$ -valued *controlled switching diffusion* process given by the stochastic differential equations

$$(2.4) \quad \begin{aligned} dX(t) &= b(X(t), S(t), v(t)) dt + \text{diag}(\sigma_1, \dots, \sigma_N) dW(t) \\ dS(t) &= \int_{\mathbb{R}} h(X(t), S(t-), v(t), z) \mathbf{p}(dt, dz), \end{aligned}$$

for $t \geq 0$, with $X(0) = X_0$ and $S(0) = S_0$, where

- (i) X_0 is a prescribed \mathbb{R}^N -valued random variable;
- (ii) S_0 is a prescribed \mathcal{S} -valued random variable;
- (iii) $W(\cdot) = [W_1(\cdot), \dots, W_N(\cdot)]^T$ is an N -dimensional standard Wiener process independent of X_0, S_0 ;
- (iv) $\mathbf{p}(dt, dz)$ is an $\mathfrak{M}(\mathbb{R}_+ \times \mathbb{R})$ -valued Poisson random measure with intensity $dt \times m(dz)$, where m is the Lebesgue measure on \mathbb{R} [25, p. 70];
- (v) $\mathbf{p}(\cdot, \cdot)$ and $W(\cdot)$ are independent;
- (vi) $v(\cdot)$ is a \mathcal{V} -valued process with measurable sample paths satisfying the following nonanticipative property: Let $\mathfrak{F}_t^v = \sigma\{v(s) : s \leq t\}$ and

$$\mathfrak{F}_{[t, \infty)}^{W, \mathbf{p}} = \sigma\{W(s) - W(t), \mathbf{p}(A, B) : A \in \mathfrak{B}([s, \infty)), B \in \mathfrak{B}(\mathbb{R}), s \geq t\}.$$

Then \mathfrak{F}_t^v and $\mathfrak{F}_{[t, \infty)}^{W, \mathbf{p}}$ are independent.

Such a process $v(\cdot)$ will be called an *admissible* (control) *policy*. If $v(\cdot)$ is a Dirac measure, i.e., $v(\cdot) = \delta_{u(\cdot)}$, where $u(\cdot)$ is a U -valued process, then it is called an admissible *nonrandomized* policy. An admissible policy $v(\cdot)$ is called *feedback* if $v(\cdot)$ is progressively measurable with respect to the natural filtration of $(X(\cdot), S(\cdot))$. A particular subclass of feedback policies is of special interest. A feedback policy $v(\cdot)$ is called a (nonhomogeneous) *Markov* policy if $v(\cdot) = \tilde{v}(\cdot, X(\cdot), S(\cdot))$ for a measurable map $\tilde{v} : \mathbb{R}_+ \times \mathbb{R}^N \times \mathcal{S} \rightarrow \mathcal{V}$. With an abuse of notation, the map \tilde{v} itself is called a Markov policy. If \tilde{v} has no explicit time dependence, it is called a *homogeneous* Markov policy. Thus, a homogeneous Markov nonrandomized policy can be identified with a measurable map $v : \mathbb{R}^N \times \mathcal{S} \rightarrow U$.

If $(W(\cdot), \mathbf{p}(\cdot, \cdot), X_0, S_0, v(\cdot))$, satisfying the above, are given on a prescribed probability space $(\Omega, \mathfrak{F}, P)$, then under our assumptions on b and λ , (2.4) will admit an almost surely unique strong solution [22, Chap. 3], [25, Chap. 3, §2c], and $X(\cdot) \in C(\mathbb{R}_+; \mathbb{R}^N)$, $S(\cdot) \in D(\mathbb{R}_+; \mathcal{S})$, where $D(\mathbb{R}_+; \mathcal{S})$ is the space of right continuous functions on \mathbb{R}_+ with left limits taking values in \mathcal{S} . However, if $v(\cdot)$ is a feedback policy, then there exists a measurable map

$$f : \mathbb{R}_+ \times C(\mathbb{R}_+; \mathbb{R}^N) \times D(\mathbb{R}_+; \mathcal{S}) \longrightarrow \mathcal{V}$$

such that for each $t \geq 0$, $v(t) = f(t, X(\cdot), S(\cdot))$ and is measurable with respect to the σ -field generated by $\{X(s), S(s) : s \leq t\}$. Thus, $v(\cdot)$ cannot be specified a priori in (2.4). Instead, we must replace $v(t)$ in (2.4) by $f(t, X(\cdot), S(\cdot))$ and (2.4) takes the form

$$\begin{aligned} dX(t) &= b(X(t), S(t), f(t, X(\cdot), S(\cdot))) dt + \text{diag}(\sigma_1, \dots, \sigma_N) dW(t), \\ (2.5) \quad dS(t) &= \int_{\mathbb{R}} h(X(t), S(t-), f(t, X(\cdot), S(\cdot)), z) \mathbf{p}(dt, dz), \end{aligned}$$

for $t \geq 0$, with $X(0) = X_0$ and $S(0) = S_0$. In general, (2.5) will not even admit a weak solution. However, if the feedback policy is a Markov policy, then the existence of a unique strong solution can be established. We now introduce some notation that will be used throughout. Define

$$L^1(\mathbb{R}^N \times \mathcal{S}) = \{f : \mathbb{R}^N \times \mathcal{S} \longrightarrow \mathbb{R} : \text{for each } i \in \mathcal{S}, f(\cdot, i) \in L^1(\mathbb{R}^N)\}.$$

$L^1(\mathbb{R}^N \times \mathcal{S})$ is endowed with the product topology of $(L^1(\mathbb{R}^N))^M$. Similarly, we define $C_0^\infty(\mathbb{R}^N \times \mathcal{S})$, $W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S})$, etc. For $f \in W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S})$ and $u \in U$, we write

$$(2.6) \quad L^u f(x, i) = L_i^u f(x, i) + \sum_{j=1}^M \bar{\lambda}_{ij}(x, u) [f(x, j) - f(x, i)],$$

where

$$(2.7) \quad L_i^u f(x, i) = \frac{1}{2} \sum_{j=1}^N \sigma_j^2 \frac{\partial^2 f(x, i)}{\partial x_j^2} + \sum_{j=1}^N \bar{b}_j(x, i, u) \frac{\partial f(x, i)}{\partial x_j},$$

and more generally, for $v \in \mathcal{V}$,

$$(2.8) \quad L^v f(x, i) = \int_U L^u f(x, i) v(du).$$

THEOREM 2.1. *Under a Markov policy v , (2.4) admits an almost surely unique strong solution such that $(X(\cdot), S(\cdot))$ is a Feller process with differential generator L^v .*

Proof (Sketch). This proof is based on the technique involving the removal of drift [41], [10, Thm. 1.4, pp. 10–12]. Clearly, it suffices to prove the result in the interval $[0, T]$, for a fixed $T > 0$. For $T > 0$, let H be the function space defined by

$$(2.9) \quad H = \left\{ g \in W_{loc}^{1,2,p}([0, T] \times \mathbb{R}^N \times \mathcal{S}), 2 \leq p < \infty : \text{ for each } i \in \mathcal{S}, \right. \\ \left. \sup_{0 \leq t \leq T} |g(t, x, i)| \text{ grows slower than } \exp(k\|x\|^2) \text{ for all } k > 0 \right\}.$$

Fix an $i \in \mathcal{S}$. For $1 \leq j \leq N$, let $\varphi_i(t, x, j)$ be the unique solution in H (as in (2.9)) of

$$(2.10) \quad \frac{\partial \varphi_i(t, x, j)}{\partial t} + L_j^{v(t, x, j)} \varphi_i(t, x, j) = 0, \\ \varphi_i(T, x, j) = x_i,$$

where $x = (x_1, \dots, x_N)$. Let $\varphi = [\varphi_1, \dots, \varphi_N]^T$. It can be shown that for fixed j , $\varphi(t, \cdot, j)$ is a homeomorphism onto its range for each $t \in [0, T]$. Set $Y(t) = \varphi(t, X(t), S(t))$, $t \in [0, T]$. Using Ito's formula, it follows that $Y(t)$ satisfies

$$(2.11) \quad Y(t) = Y(0) + \int_0^t \left[(D\varphi_s \text{diag}(\sigma_1, \dots, \sigma_N)) \circ \varphi_s^{-1} \right] (Y(s)) dW(s) \\ + \int_0^t \int_{\mathbb{R}} \left[\varphi_s \left(\varphi_s^{-1}(Y(s-)) + \tilde{h}(\varphi_s^{-1}(Y(s-)), z) \right) - Y(s) \right] \mathfrak{p}(ds, dz),$$

where $D\varphi_s$, φ_s^{-1} denote, respectively, the Jacobian matrix and the inverse map of $\varphi(s, \cdot, S(s))$, “ \circ ” indicates composition of functions and

$$\tilde{h}(\cdot, \cdot, \cdot) = [0, 0, \dots, 0, h(\cdot, \cdot, \cdot)]^T \in \mathbb{R}^{N+1}.$$

Now by [41], (2.11) has an almost surely unique strong solution, which is a Markov process. The corresponding claim for $(X(t), S(t))$ follows via the homeomorphic property of φ . It remains to show the strong Feller property. Pick any bounded continuous function $f : \mathbb{R}^N \times \mathcal{S} \rightarrow \mathbb{R}$. The system of equations

$$(2.12) \quad \frac{\partial \psi(t, x, i)}{\partial t} + L^{v(t, x, i)} \psi(t, x, i) = 0, \\ \psi(T, x, i) = f(x, i),$$

can be shown to have a unique solution in H [18]. Therefore, by Ito's formula, it follows that

$$\psi(t, x, i) = E[f(X(T), S(T)) \mid X(t) = x, S(t) = i],$$

where the expectation is under the Markov policy v . By Sobolev's imbedding theorem [5, p. 53], $H \subset C([0, T] \times \mathbb{R}^N \times \mathcal{S})$ and hence $\psi(t, \cdot, i)$ is continuous for each $t \in [0, T]$. \square

Some comments are in order now.

Remark 2.1. (i) We have used Ito's formula for functions in $W_{loc}^{1,2,p}(\mathbb{R}_+ \times \mathbb{R}^N \times \mathcal{S})$. This generalization is due to Krylov [28, pp. 121–127] for “classical” diffusions. Its extension for the present system is routine.

(ii) The wellposedness of the Cauchy problem for the weakly coupled parabolic system (2.10) has been established in [18] under slightly stronger conditions on the first-order terms. However, in view of the results in [3] and [30, Chap. 7], its extension to the present case is straightforward.

Remark 2.2. We have seen in Theorem 2.1 that under a Markov policy the corresponding solution $(X(\cdot), S(\cdot))$ of (2.4) is a Markov process. We have the following converse result. Let $v(\cdot)$ be a feedback policy, such that the corresponding solution $(X(\cdot), S(\cdot))$ of (2.4) is a Markov process. Then $v(\cdot)$ may be taken to be a Markov policy. Since we do not need this result, we omit the proof.

2.1. The optimization problem. Let $\bar{c} : \mathbb{R}^N \times \mathcal{S} \times U \rightarrow \mathbb{R}_+$ be a bounded, continuous cost function, and let $c : \mathbb{R}^N \times \mathcal{S} \times \mathcal{V} \rightarrow \mathbb{R}_+$ be defined as

$$c(\cdot, \cdot, v) = \int_U \bar{c}(\cdot, \cdot, u) v(du).$$

Let $\alpha > 0$ be a prescribed discount factor. Let $v(\cdot)$ be an admissible policy and $(X(\cdot), S(\cdot))$ the corresponding process. Then the total α -discounted cost under $v(\cdot)$ is defined as

$$(2.13) \quad J_v(x, i) := E \left[\int_0^\infty e^{-\alpha t} c(X(t), S(t), v(t)) dt \mid X(0) = x, S(0) = i \right].$$

If the laws of X_0, S_0 are $\pi \in \mathcal{P}(\mathbb{R}^N)$, $\xi \in \mathcal{P}(\mathcal{S})$, respectively, then

$$(2.14) \quad J_v(\pi, \xi) = \sum_i \int_{\mathbb{R}^N} J_v(x, i) \pi(dx) \xi(i).$$

Let

$$(2.15) \quad V(x, i) := \inf_{v(\cdot)} \{ J_v(x, i) \},$$

$$(2.16) \quad V(\pi, \xi) := \inf_{v(\cdot)} \{ J_v(\pi, \xi) \}.$$

The function $V(x, i)$ is called the (α -discounted) value function. An admissible policy $v(\cdot)$ satisfying

$$J_v(\pi, \xi) = V(\pi, \xi)$$

is called an optimal policy for the initial law (π, ξ) . An admissible policy is called optimal if it is optimal for any initial law. Our aim is to find an admissible optimal policy which is homogeneous Markov and nonrandomized.

We now introduce the (discounted) occupation measures [12]. Let $v(\cdot)$ be an admissible policy and $(X(\cdot), S(\cdot))$ the corresponding process with initial law (π, ξ) . Define the occupation measure $\nu[\pi, \xi; v] \in \mathcal{P}(\mathbb{R}^N \times \mathcal{S} \times U)$ by

$$(2.17) \quad \int f d\nu[\pi, \xi; v] = \alpha E \left[\int_0^\infty e^{-\alpha t} \int_U f(X(t), S(t), u) v(t)(du) dt \right]$$

for $f \in C_b(\mathbb{R}^N \times \mathcal{S} \times U)$. Also, we define

$$(2.18) \quad M_1[\pi, \xi] = \{\nu[\pi, \xi; v] : v(\cdot) \text{ is admissible}\},$$

$$(2.19) \quad M_2[\pi, \xi] = \{\nu[\pi, \xi; v] : v(\cdot) \text{ is homogeneous Markov}\},$$

$$(2.20) \quad M_3[\pi, \xi] = \{\nu[\pi, \xi; v] : v(\cdot) \text{ is homogeneous nonrandomized Markov}\}.$$

In terms of these occupation measures

$$(2.21) \quad J_v(\pi, \xi) = \alpha^{-1} \int \bar{c} d\nu[\pi, \xi; v].$$

We will show in §4 that $M_1[\pi, \xi] = M_2[\pi, \xi]$ and that $M_2[\pi, \xi]$ is compact, convex, and $M_2^e[\pi, \xi] \subset M_3[\pi, \xi]$, where $M_2^e[\pi, \xi]$ is the set of extreme points of $M_2[\pi, \xi]$. Thus, for a fixed initial law, the optimization problem (2.13) will reduce to a convex optimization problem in view of (2.21).

3. Compactness of laws. We will establish the compactness of laws of the process $(X(\cdot), S(\cdot))$ under various policies using the approach in [11, Chap. 2]. Let $\pi_0 \in \mathcal{P}(\mathbb{R}^N)$, $\xi_0 \in \mathcal{P}(\mathcal{S})$. Let $\mathcal{L}_i[\pi_0, \xi_0] \subset \mathcal{P}(C(\mathbb{R}_+; \mathbb{R}^N) \times D(\mathbb{R}_+; \mathcal{S}))$, $i = 1, 2, 3$, denote the set of laws of $(X(\cdot), S(\cdot))$ under all admissible/Markov/homogeneous Markov policies with fixed initial law (π_0, ξ_0) .

THEOREM 3.1. *The set $\mathcal{L}_1[\pi_0, \xi_0]$ is compact in $\mathcal{P}(C(\mathbb{R}_+; \mathbb{R}^N) \times D(\mathbb{R}_+; \mathcal{S}))$.*

Proof. It clearly suffices to replace \mathbb{R}_+ by $[0, T]$ for arbitrary $T > 0$. Fix $T > 0$. Let $(X^n(\cdot), S^n(\cdot), W^n(\cdot), \mathbf{p}^n(\cdot, \cdot), v^n(\cdot), X_0^n, S_0^n)$, $n \geq 1$, satisfy (2.4) on probability spaces $(\Omega^n, \mathfrak{F}^n, P^n)$ respectively, the laws of X_0^n, S_0^n being π_0, ξ_0 respectively for all n . Let $\{f_i\}$ be a countable dense subset of the unit ball of $C(U)$. Define $\beta_j^n(t) = \int f_j dv^n(t)$, $t \in [0, T]$. Let B denote closed unit ball of $L^\infty[0, T]$ with the topology given by the weak topology of $L^2[0, T]$ relative to B . Let E be a countable product of replicas of B . Since B is compact and metrizable and hence Polish, the same follows for E . Let $\beta^n(\cdot) = [\beta_1^n(\cdot), \beta_2^n(\cdot), \dots]$, $n \geq 1$, viewed as E -valued random variables. Using the assumed conditions on b , it can be easily shown that for $t_1, t_2 \in [0, T]$,

$$E\left[\|X^n(t_2) - X^n(t_1)\|^4\right] \leq K|t_2 - t_1|^2$$

for some T -dependent $K > 0$. It follows that the laws of the sequence $\{X^n(\cdot)\}$ are tight in $\mathcal{P}(C(\mathbb{R}_+; \mathbb{R}^N))$. Since \mathcal{S} is finite and E is compact, it follows by Prohorov's theorem [24, Thm. 2.6, p. 7] that, for $A_1 \in \mathfrak{B}(\mathbb{R}_+)$, $A_2 \in \mathfrak{B}(\mathbb{R})$ fixed, the sequence $(X^n(\cdot), S^n(\cdot), \beta^n(\cdot), W^n(\cdot), \mathbf{p}^n(A_1 \times A_2))$ converges to a limit

$$(X(\cdot), S(\cdot), \beta(\cdot), W(\cdot), \mathbf{p}(A_1 \times A_2)).$$

Dropping to a subsequence if necessary and invoking Skorohod's theorem [24, p. 9], we may assume that all these random variables are defined on a common probability space and the convergence is almost surely on this probability space. By [11, Lemma II.1.2, p. 24] we can find a \mathcal{V} -valued process $v(\cdot)$ such that $\beta_i(t) = \int f_i dv(t)$, $i \geq 1$. Define $Z^n(\cdot) = [Z_1^n(\cdot), \dots, Z_N^n(\cdot)]^T$, $Y^n(\cdot)$, $n \geq 1$, by

$$Z_i^n(t) = X_i^n(t) - \int_0^t b_i(X^n(s), S^n(s), v^n(s)) ds, \quad t \geq 0,$$

$$Y^n(t) = S^n(t) - \sum_{j=1}^M \int_0^t \lambda_{S^n(s-), j}(X^n(s), v^n(s)) (j - S^n(s-)) ds, \quad t \geq 0,$$

and $Z(\cdot) = [Z_1(\cdot), \dots, Z_N(\cdot)]^T$, $Y(\cdot)$ by

$$Z_i(t) = X_i(t) - \int_0^t b_i(X(s), S(s), v(s)) ds, \quad t \geq 0,$$

$$Y(t) = S(t) - \sum_{j=1}^M \int_0^t \lambda_{S(s-), j}(X(s), v(s)) (j - S(s-)) ds, \quad t \geq 0.$$

Then, by [11, Lemma II.1.3, p. 26] and standard representation theorems for semimartingales [25, pp. 172–178] applied to $Z_i(t)$ and $Y(t)$, it follows that on an augmented probability space $(X(\cdot), S(\cdot))$ satisfies (2.4) for an admissible policy $v(\cdot)$ and driven by a Wiener process $\bar{W}(\cdot)$ and a Poisson random measure $\tilde{p}(\cdot, \cdot)$. \square

We now state the next theorem without proof as it would be almost identical to the proof in [11, Thm. II.2.1, p. 29], in view of the estimates in [30, p. 582].

THEOREM 3.2. *The sets $\mathcal{L}_2[\pi_0, \xi_0]$, $\mathcal{L}_3[\pi_0, \xi_0]$ are compact.*

Let $\{v_n\}$ be a sequence of homogeneous Markov policies and $(X^n(\cdot), S^n(\cdot))$ the corresponding solutions of (2.4) with $X^n(0) = x_0$, $S^n(0) = i_0$ for all $n \geq 0$. Let $p^n(t, x_0, i_0, y, j)$ be the fundamental solutions corresponding to the operators $(\frac{\partial}{\partial t} + L^{v_n})$. Let $(X^n(\cdot), S^n(\cdot)) \rightarrow (X^\infty(\cdot), S^\infty(\cdot))$, where the latter is governed by a homogeneous Markov policy v_∞ . Then, using the Hölder estimates on $p^n(t, x_0, i_0, y, j)$ [30, p. 582], we can show the following result as in [10, Thm. II.2.2, p. 33].

LEMMA 3.1. *For each $t > 0$, $p^n(t, x_0, i_0, \cdot, \cdot) \rightarrow p^\infty(t, x_0, i_0, \cdot, \cdot)$ in $L^1(\mathbb{R}^N \times \mathcal{S})$. In other words, the laws of $(X^n(t), S^n(t))$ converge to that of $(X^\infty(t), S^\infty(t))$ in total variation.*

Next, we introduce a topology to the space of all homogeneous Markov policies. Let $F = \{v : \mathbb{R}^N \times \mathcal{S} \rightarrow \mathcal{V} : v \text{ is measurable}\}$. We endow F with the topology described in [11, p. 30]. Then F is a compact metric space. Its topology is determined by the following convergence criterion [11, Lemma II.2.1, p. 32].

LEMMA 3.2. *Let $f \in L^2(\mathbb{R}^N \times \mathcal{S}) \cap L^1(\mathbb{R}^N \times \mathcal{S})$, $g \in C_b(\mathbb{R}^N \times \mathcal{S} \times U)$ and $v_n \rightarrow v$ in F . Then*

$$(3.1) \quad \int_{\mathbb{R}^N} f(x, i) \int_{\mathcal{S}} g(x, i, \cdot) dv_n(x, i) dx \xrightarrow{n \rightarrow \infty} \int_{\mathbb{R}^N} f(x, i) \int_{\mathcal{S}} g(x, i, \cdot) dv(x, i) dx$$

for each $i \in \mathcal{S}$. Conversely, if (3.1) holds for all such f, g , and $i \in \mathcal{S}$, then $v_n \rightarrow v$ in F .

Let $\mathcal{L}(v)$ denote the law of $(X(\cdot), S(\cdot))$ when $X(0) = x_0$, $S(0) = i_0$ and the homogeneous Markov policy v is used. Using Lemma 3.1, the following theorem can be proved exactly the same way as in [11, Thm. II.2.3, p. 34].

THEOREM 3.3. *The map $v \mapsto \mathcal{L}(v)$ from F into $\mathcal{P}(C(\mathbb{R}_+; \mathbb{R}^N) \times D(\mathbb{R}_+; \mathcal{S}))$ is continuous.*

4. Convexity and extremality of occupation measures. In this section we will study the properties of the occupation measures $\nu[\pi, \xi; v]$ introduced in (2.17), following the approach in [12].

LEMMA 4.1. *The sets $M_1[\pi, \xi]$, $M_2[\pi, \xi]$, $M_3[\pi, \xi]$ as defined in (2.18)–(2.20) are compact.*

Proof. This follows from Theorems 3.1 and 3.2. \square

LEMMA 4.2. *For each fixed initial law (π, ξ) , $M_1[\pi, \xi] = M_2[\pi, \xi]$.*

Proof. Let $\nu[\pi, \xi; v] \in M_1$. Disintegrate it as

$$(4.1) \quad \nu[\pi, \xi; v](dx \times \{i\} \times du) = \bar{\nu}[\pi, \xi; v](dx \times \{i\}) \bar{\nu}(x, i)(du),$$

where $\bar{\nu}[\pi, \xi; v]$ is the marginal of $\nu[\pi, \xi; v]$ on $\mathbb{R}^N \times \mathcal{S}$ and $\bar{v}(x, i)$ is a version of the regular conditional law defined $\bar{\nu}[\pi, \xi; v]$ almost surely. Pick any version from this equivalence class and keep it fixed henceforth. The map $\bar{v}(\cdot, \cdot)$ obviously defines a homogeneous Markov policy. Let $(X'(\cdot), S'(\cdot))$ be the solution of (2.4) with $v(\cdot)$ replaced by $v'(\cdot) = \bar{v}(X'(\cdot), S'(\cdot))$ and with initial law (π, ξ) . Let $f \in C_b(\mathbb{R}^N \times \mathcal{S} \times U)$ and let

$$(4.2) \quad \varphi(x, i) = E \left[\int_0^\infty e^{-\alpha t} \int_U f(X'(t), S'(t), u) v'(t)(du) dt \mid X'(0) = x, S'(0) = i \right].$$

Using the strong Markov property of $(X'(\cdot), S'(\cdot))$ (this follows from the Feller property) and the local solvability of weakly coupled systems of elliptic equations [31, Chap. 7, p. 388], it can be shown by employing standard arguments involving Ito's formula that $\varphi(x, i)$ is the unique solution in $W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S}) \cap C_b(\mathbb{R}^N \times \mathcal{S})$, $2 \leq p < \infty$, to

$$(4.3) \quad L^{\bar{v}(x,i)} \varphi(x, i) - \alpha \varphi(x, i) + \int_U f(x, i, u) \bar{v}(x, i)(du) = 0.$$

Define a process $Y(\cdot)$ by

$$Y(t) = \int_0^t \int_U e^{-\alpha s} f(X(s), S(s), u) v(s)(du) ds + e^{-\alpha t} \varphi(X(t), S(t)).$$

Then

$$(4.4) \quad \begin{aligned} E[Y(t)] - E[Y(0)] &= E[Y(t)] - \sum_{j=1}^M \int_{\mathbb{R}^N} \varphi(x, j) \pi(dx) \xi(j) \\ &= E \left[\int_0^t e^{-\alpha s} \left[L^{v(s)} \varphi(X(s), S(s)) - \alpha \varphi(X(s), S(s)) \right. \right. \\ &\quad \left. \left. + \int_U f(X(s), S(s), u) v(s)(du) \right] ds \right]. \end{aligned}$$

Letting $t \rightarrow \infty$ and using the definition of $\bar{v}(\cdot, \cdot)$ and (4.3), it follows that the right-hand side in (4.4) tends to zero (cf. [11, Thm. 4.2, pp. 40–42]). Thus,

$$\begin{aligned} \lim_{t \rightarrow \infty} E[Y(t)] &= E[\varphi(X_0, S_0)] \\ &= E[\varphi(X'_0, S'_0)]. \end{aligned}$$

Since $f \in C_b(\mathbb{R}^N \times \mathcal{S} \times U)$ was arbitrary, it follows that $\nu[\pi, \xi; v] = \nu[\pi, \xi; \bar{v}]$. \square

Let $\nu[\pi, \xi; v] \in M_2[\pi, \xi]$. By a routine extension of the inequality [28, p. 66], it follows that $\bar{\nu}[\pi, \xi; v]$ (as in (4.1)) is absolutely continuous with respect to the product of the Lebesgue measure on \mathbb{R}^N and the counting measure on \mathcal{S} and therefore has a density $\varphi[\pi, \xi; v]$. Let $\bar{\nu}[\pi, \xi; v]$ be the marginal of $\bar{\nu}[\pi, \xi; v]$ on \mathcal{S} . With “supp” denoting the support of a measure, let

$$(4.5) \quad \text{supp}(\bar{\nu}[\pi, \xi; v]) = S_1[\pi, \xi; v] \subset \mathcal{S}.$$

It is not difficult to see that $\varphi[\pi, \xi; v](x, i) > 0$ almost everywhere $x \in \mathbb{R}^N$, $i \in S_1[\pi, \xi; v]$ and $\varphi[\pi, \xi; v](x, i) = 0$ for $i \in \mathcal{S} \setminus S_1[\pi, \xi; v]$. For $f \in W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S})$ define

$$(4.6) \quad L_\alpha^v f(x, i) = L^{v(x,i)} f(x, i) - \alpha f(x, i).$$

Then, $\varphi[\pi, \xi; v]$ is the unique solution in $L^1(\mathbb{R}^N \times \mathcal{S})$ to

$$(4.7) \quad \begin{aligned} \sum_{i=1}^M \int_{\mathbb{R}^N} L_{\alpha}^{v(x,i)} g(x, i) \varphi(x, i) dx &= - \sum_{i=1}^M \int_{\mathbb{R}^N} g(x, i) \pi(dx) \xi(i), \\ \sum_{i=1}^M \int_{\mathbb{R}^N} \varphi(x, i) dx &= 1, \quad \varphi(x, i) \geq 0, \end{aligned}$$

for every $g \in C_0^\infty(\mathbb{R}^N \times \mathcal{S})$. Using the above, we will show that $M_2[\pi, \xi]$ is convex.

LEMMA 4.3. *The set $M_2[\pi, \xi]$ is convex.*

Proof. Let v_1, v_2 be two homogeneous Markov policies and $0 \leq a \leq 1$. Define a homogeneous Markov policy by

$$(4.8) \quad v(x, i) = \frac{a\varphi[\pi, \xi; v_1](x, i)v_1(x, i) + (1-a)\varphi[\pi, \xi; v_2](x, i)v_2(x, i)}{a\varphi[\pi, \xi; v_1](x, i) + (1-a)\varphi[\pi, \xi; v_2](x, i)}$$

for $(x, i) \in \mathbb{R}^N \times \{S_1[\pi, \xi; v_1] \cup S_1[\pi, \xi; v_2]\}$ and arbitrary otherwise. Let $f \in C_0^\infty(\mathbb{R}^N \times \mathcal{S})$. It is easy to see that

$$L_{\alpha}^{v(x,i)} f(x, i) = \frac{a\varphi[\pi, \xi; v_1](x, i)L_{\alpha}^{v_1(x,i)} f(x, i) + (1-a)\varphi[\pi, \xi; v_2](x, i)L_{\alpha}^{v_2(x,i)} f(x, i)}{a\varphi[\pi, \xi; v_1](x, i) + (1-a)\varphi[\pi, \xi; v_2](x, i)}.$$

Let $\varphi(x, i) = a\varphi[\pi, \xi; v_1](x, i) + (1-a)\varphi[\pi, \xi; v_2](x, i)$. From (4.7) and (4.8) it follows that $\varphi = \varphi[\pi, \xi; v]$. Thus,

$$\begin{aligned} \nu[\pi, \xi; v](dx \times \{i\} \times du) &= \varphi[\pi, \xi; v](x, i) dx v(x, i)(du) \\ &= a\varphi[\pi, \xi; v_1](x, i) dx v_1(x, i)(du) + (1-a)\varphi[\pi, \xi; v_2](x, i) dx v_2(x, i)(du) \\ &= (a\nu[\pi, \xi; v_1] + (1-a)\nu[\pi, \xi; v_2])(dx \times \{i\} \times du). \quad \square \end{aligned}$$

Let

$$(4.9) \quad \mathcal{I}[\pi, \xi] = \{\bar{\nu}[\pi, \xi; v] \in \mathcal{P}(\mathbb{R}^N \times \mathcal{S}) : \nu[\pi, \xi; v] \in M_2[\pi, \xi]\},$$

where $\bar{\nu}[\pi, \xi; v]$ is as in (4.1).

The proof of the next lemma is analogous to that of [12, Lemma 3.2]. We present a brief sketch describing the essential ideas.

LEMMA 4.4. *The set $\mathcal{I}[\pi, \xi]$ is compact in $\mathcal{P}(\mathbb{R}^N \times \mathcal{S})$ in total variation.*

Proof. By a routine extension of the inequality [28, p. 66] to the present case, $\varphi[\pi, \xi; v]$ will be uniformly bounded in $L^p(\mathbb{R}^N)$. For the sake of convenience, assume that the initial condition is $(x_0, i_0) \in \mathbb{R}^N \times \mathcal{S}$. As in [11, Lemma 5.2, p. 44], we can show by considering appropriate estimates on the weakly coupled systems of elliptic equations [31, Chap. 7] that for any bounded open set A such that $\bar{A} \subset \mathbb{R}^N \setminus \{x_0\}$ and $i \in \mathcal{S} \setminus \{i_0\}$ there exists a $\beta > 0$ and a $K \in (0, \infty)$ such that

$$(4.10) \quad |\varphi[x_0, i_0; v](y, i) - \varphi[x_0, i_0; v](z, i)| \leq K\|y - z\|^\beta, \quad y, z \in A,$$

under any choice of a homogeneous Markov policy v . By Theorem 3.2, $\mathcal{I}[x_0, i_0]$ is compact in the Prohorov topology of $\mathcal{P}(\mathbb{R}^N \times \mathcal{S})$. Let $\{\bar{\nu}[x_0, i_0; v_n]\}$ be a sequence in

$\mathcal{I}[x_0, i_0]$ and $\bar{\nu}[x_0, i_0; v_\infty]$ be a weak limit point of $\{\bar{\nu}[x_0, i_0; v_n]\}$. We need to show that $\varphi[x_0, i_0; v_n] \rightarrow \varphi[x_0, i_0; v_\infty]$ in $L^1(\mathbb{R}^N \times \mathcal{S})$. The equicontinuity of $\{\varphi[x_0, i_0; v_n]\}$ follows from (4.10). Also (4.10) together with the uniform L^p -estimates implies pointwise boundedness. Thus, by the Arzela–Ascoli theorem, we may drop to a subsequence, if necessary, to conclude that for each $i \in \mathcal{S}$,

$$\varphi[x_0, i_0, v_n](\cdot, i) \rightarrow \psi(\cdot, i),$$

for some $\psi(\cdot, i)$, uniformly on compact subsets of \mathbb{R}^N . By the uniform L^p -estimates the convergence is also in $L^1(\mathbb{R}^N)$. Thus

$$(4.11) \quad \int_{\mathbb{R}^N} \varphi[x_0, i_0; v_n](y, i) f(y) dy \xrightarrow{n \rightarrow \infty} \int_{\mathbb{R}^N} \psi(y, i) f(y) dy,$$

for all $f \in C_b(\mathbb{R}^N)$. But (4.11) certainly holds with $\varphi[x_0, i_0; v_\infty](\cdot, i)$ replacing $\psi(\cdot, i)$. Therefore, $\varphi[x_0, i_0; v_\infty] \equiv \psi$. \square

We are now in a position to characterize the extreme points of $M_2[\pi, \xi]$. Let v be a homogeneous Markov policy such that, for each $x \in \mathbb{R}^N$ and $i \in \mathcal{S}$,

$$(4.12) \quad v(x, i) = av_1(x, i) + (1 - a)v_2(x, i),$$

where $a \in (0, 1)$ and v_1, v_2 are distinct homogeneous Markov policies, i.e., there exists at least one $i_0 \in \mathcal{S}$ such that $v_1(\cdot, i_0)$ and $v_2(\cdot, i_0)$ differ on a set of strictly positive measure. The proof of the next lemma closely follows that of [12, Lemma 3.3]; we therefore present only a brief sketch of the proof.

LEMMA 4.5. *Let v be as in (4.12). Then, $\nu[\pi, \xi; v]$ is not an extreme point of $M_2[\pi, \xi]$.*

Proof. We will show that if v satisfies (4.12), then there are homogeneous Markov policies \tilde{v}_1, \tilde{v}_2 , and $b \in (0, 1)$ such that

$$\nu[\pi, \xi; v] = b\nu[\pi, \xi; \tilde{v}_1] + (1 - b)\nu[\pi, \xi; \tilde{v}_2].$$

It suffices to find $b \in (0, 1)$ and \tilde{v}_1, \tilde{v}_2 satisfying

$$(4.13) \quad v(x, i) = \frac{b\varphi[\pi, \xi; \tilde{v}_1](x, i)\tilde{v}_1(x, i) + (1 - b)\varphi[\pi, \xi; \tilde{v}_2](x, i)\tilde{v}_2(x, i)}{b\varphi[\pi, \xi; \tilde{v}_1](x, i) + (1 - b)\varphi[\pi, \xi; \tilde{v}_2](x, i)}$$

for $(x, i) \in \mathbb{R}^N \times \{S_1[\pi, \xi; \tilde{v}_1] \cup S_1[\pi, \xi; \tilde{v}_2]\}$ (see (4.5)). For $R > 0$, let v'_1, v'_2 be homogeneous Markov policies defined by

$$(4.14) \quad v'_j(x, i) = \begin{cases} v_j(x, i), & \|x\| \leq R \\ v(x, i), & \|x\| > R \end{cases} \quad i \in \mathcal{S}, \quad j = 1, 2.$$

Let $\bar{v}(\cdot)$ be a given homogeneous Markov policy. Define a homogeneous Markov policy v''_2 via

$$(4.15) \quad \begin{aligned} v(x, i) &= av'_1(x, i) + (1 - a)v'_2(x, i) \\ &= \frac{b\varphi[\pi, \xi; v'_1](x, i)v'_1(x, i) + (1 - b)\varphi[\pi, \xi; \bar{v}](x, i)v''_2(x, i)}{b\varphi[\pi, \xi; v'_1](x, i) + (1 - b)\varphi[\pi, \xi; \bar{v}](x, i)} \end{aligned}$$

for $(x, i) \in \mathbb{R}^N \times \{S_1[\pi, \xi; v'_1] \cup S_1[\pi, \xi; \bar{v}]\}$ and arbitrary otherwise. The arguments used in the proof of [12, Lemma 3.3] mutatis mutandis will ensure a suitable choice of $b \in (0, 1)$ such that v''_2 is a genuine homogeneous Markov policy. Fix a $b \in (0, 1)$ as in (4.15). Given a homogeneous Markov policy $\bar{v}(\cdot)$, we obtain $v''_2(\cdot)$ via (4.15). Thus, we have a map $\bar{v}[\pi, \xi; \bar{v}] \mapsto \bar{v}[\pi, \xi; v''_2]$ from $\mathcal{I}[\pi, \xi]$ to $\mathcal{I}[\pi, \xi]$. Using Lemma 4.4, it can be shown as in the proof of [11, Lemma 3.3] that this map is continuous in the total variation. By Schauder's fixed point theorem [26, p. 220], this map has a fixed point. In other words, there exists a homogeneous Markov policy v''_2 such that

$$v(x, i) = \frac{b\varphi[\pi, \xi; v'_1](x, i)v'_1(x, i) + (1 - b)\varphi[\pi, \xi; v''_2]v''_2(x, i)}{b\varphi[\pi, \xi; v'_1](x, i) + (1 - b)\varphi[\pi, \xi; v''_2](x, i)}$$

for $(x, i) \in \mathbb{R}^N \times \{S_1[\pi, \xi; v'_1] \cup S_1[\pi, \xi; v''_2]\}$. Since $v'_1 \neq v$ on a set of strictly positive measure for sufficiently large R , $v''_2 \neq v'_1$ on this set. Thus

$$\nu[\pi, \xi; v] = b\nu[\pi, \xi; v'_1] + (1 - b)\nu[\pi, \xi; v''_2]$$

as desired. \square

The results in this section are now summarized as follows.

THEOREM 4.1. $M_1[\pi, \xi] = M_2[\pi, \xi]$, and $M_2[\pi, \xi]$ is compact and convex, and each of its extreme points corresponds to some $\nu[\pi, \xi; v]$, where v is a homogeneous Markov nonrandomized policy.

5. Existence of an optimal policy. Using the results of the previous section, we will establish the existence of an optimal policy.

THEOREM 5.1. *There exists a homogeneous Markov optimal policy.*

Proof. Let $(\pi, \xi) \in \mathcal{P}(\mathbb{R}^N) \times \mathcal{P}(\mathcal{S})$ such that $\text{supp}(\pi) = \mathbb{R}^N$ and $\text{supp}(\xi) = S$. Since \bar{c} is bounded and continuous, the map $M_2[\pi, \xi] \ni \nu \mapsto \int \bar{c} d\nu$ is continuous. Thus, there exists a homogeneous Markov policy v^* such that

$$J_{v^*}(\pi, \xi) = \min_v \{J_v(\pi, \xi) : v \text{ is homogeneous Markov}\}.$$

By Lemma 4.2, it follows that

$$J_{v^*}(\pi, \xi) = V(\pi, \xi).$$

Therefore, v^* is optimal for the initial law (π, ξ) . We will show that v^* is optimal for any initial law. It suffices to show that v^* is optimal for any initial condition $(x, i) \in \mathbb{R}^N \times \mathcal{S}$. Suppose there exist $(x_0, i_0) \in \mathbb{R}^N \times \mathcal{S}$ and a homogeneous Markov policy v such that

$$(5.1) \quad J_v(x_0, i_0) < J_{v^*}(x_0, i_0).$$

Using the fact that the solution of (2.4) under a Markov policy is a Feller process, it can be easily shown that the function $J_v(x, i)$ is continuous in x for each v . Thus, (5.1) holds in a neighborhood B of x_0 . Define a policy v' by

$$v'(t) = v^*(X(t), S(t))I\{X_0 \notin B\} + v(X(t), S(t))I\{X_0 \in B\},$$

where $(X(\cdot), S(\cdot))$ is governed by $v'(\cdot)$. Then, it is easily shown that

$$J_{v'}(\pi, \xi) < J_{v^*}(\pi, \xi),$$

which is a contradiction. Thus, v^* is optimal. \square

THEOREM 5.2. *There exists a homogeneous Markov nonrandomized optimal policy.*

Proof. Let v^* be as in Theorem 5.1. Let $M_2^e[\pi, \xi]$ be the set of extreme points of $M_2[\pi, \xi]$. Since $M_2[\pi, \xi]$ is compact, by Choquet's theorem [34], $\nu[\pi, \xi; v^*]$ is the barycenter of a probability measure m supported on $M_2^e[\pi, \xi]$. Therefore,

$$(5.2) \quad \int \bar{c} d\nu[\pi, \xi; v^*] = \int_{M_2^e[\pi, \xi]} \left(\int \bar{c} d\mu \right) m(d\mu).$$

Since v^* is optimal, it follows from (5.2) that there exists a $\nu[\pi, \xi; v] \in M_2^e[\pi, \xi]$ such that

$$\int \bar{c} d\nu[\pi, \xi; v^*] = \int \bar{c} d\nu[\pi, \xi; v].$$

Thus, v is also optimal. By Theorem 4.1 it is nonrandomized. \square

6. Dynamic programming equations. Using the existence results of the previous section, we will now derive the dynamic programming or Hamilton–Jacobi–Bellman (HJB) equations which in our case will be a weakly coupled system of quasi-linear elliptic equations, and then characterize the optimal policy as a minimizing selector of an appropriate “Hamiltonian.” The HJB equations for our problem are

$$(6.1) \quad \alpha\psi(x, i) = \inf_{u \in U} \{L^u\psi(x, i) + \bar{c}(x, i, u)\}.$$

THEOREM 6.1. *The value function $V(x, i)$ is the unique solution of (6.1) in the space $W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S}) \cap C_b(\mathbb{R}^N \times \mathcal{S})$ for any $2 \leq p < \infty$.*

Proof. We have already seen in the proof of Theorem 5.1 that $V(x, i) \in C_b(\mathbb{R}^N \times \mathcal{S})$. Let v^* be a homogeneous Markov nonrandomized optimal policy and $(X(\cdot), S(\cdot))$ the corresponding solution of (2.4). Then, for $(x, i) \in \mathbb{R}^N \times \mathcal{S}$,

$$(6.2) \quad V(x, i) = E \left[\int_0^\infty e^{-\alpha t} \bar{c}(X(t), S(t), v^*(X(t), S(t))) dt \mid X(0) = x, S(0) = i \right].$$

By standard arguments (see the arguments following (4.2)), $V(x, i)$ is the unique solution in $W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S}) \cap C_b(\mathbb{R}^N \times \mathcal{S})$, for any $2 \leq p < \infty$, of

$$(6.3) \quad \alpha V(x, i) = L^{v^*(x, i)} V(x, i) + \bar{c}(x, i, v^*(x, i)).$$

Suppose there exist $x_0 \in \mathbb{R}^N$, $i_0 \in \mathcal{S}$, $u \in U$, and $\delta > 0$ such that

$$\alpha V(x_0, i_0) > L^u V(x_0, i_0) + \bar{c}(x_0, i_0, u) + \delta.$$

Then, by the continuity of $V(\cdot, i_0)$, the above will hold in a neighborhood $N(x_0)$ of x_0 . Define a homogeneous Markov nonrandomized policy \tilde{v} as follows:

$$\tilde{v}(x, i) = \begin{cases} v^*(x, i) & \text{if } (x, i) \notin N(x_0) \times \mathcal{S}, \\ u & \text{if } (x, i) \in N(x_0) \times \mathcal{S}. \end{cases}$$

Then

$$\alpha V(x, i_0) > L^{\tilde{v}(x, i_0)} V(x, i_0) + \bar{c}(x, i_0, \tilde{v}(x, i_0)) + \delta I\{x \in N(x_0)\}.$$

It is easily seen that

$$V(x, i_0) \geq J_{\tilde{v}}(x, i_0) + \delta'$$

for some $\delta' > 0$, which is a contradiction. Hence, $V(x, i)$ satisfies (6.1). Let V' be another solution of (6.1) in the desired class. Then it can be shown using standard arguments (cf. [11, Thm. III.2.4, pp. 69–70]) that

$$|V(x, i) - V'(x, i)| \leq 2Ke^{-\alpha t},$$

where $K > 0$ is a constant. Letting $t \rightarrow \infty$, $V \equiv V'$. \square

COROLLARY 6.1. Assume that for each $i \in \mathcal{S}$, $\bar{c}(\cdot, i, \cdot)$ is Lipschitz in its first argument uniformly with respect to the third. Then $V(x, i)$ is the unique solution of (6.1) in $C^2(\mathbb{R}^N \times \mathcal{S}) \cap C_b(\mathbb{R}^N \times \mathcal{S})$.

Proof. It suffices to show that V is C^2 . Since $V(x, i) \in W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S})$ for any $2 \leq p < \infty$, by Sobolev's imbedding theorem, $V(x, i) \in C^{1,\gamma}(\mathbb{R}^N \times \mathcal{S})$, for $0 < \gamma < 1$, γ arbitrarily close to 1, and hence by our assumptions on \bar{b} , $\bar{\lambda}$, \bar{c} , it is easy to see that

$$\alpha V(x, i) - \inf_{u \in U} \left\{ \sum_{j=1}^N \bar{b}_j(x, i, u) \frac{\partial V(x, i)}{\partial x_j} + \sum_{j=1}^M \bar{\lambda}_{ij}(x, u) (V(x, j) - V(x, i)) + \bar{c}(x, i, u) \right\}$$

is in $C^{0,\gamma}$. By elliptic regularity [23, p. 287] applied to (6.1) (V replacing ψ), we conclude that $V \in C^{2,\gamma}$. \square

THEOREM 6.2. A homogeneous Markov nonrandomized policy v is optimal if and only if

$$\begin{aligned} (6.4) \quad & \sum_{j=1}^N \bar{b}_j(x, i, v(x, i)) \frac{\partial V(x, i)}{\partial x_j} \\ & + \sum_{k=1}^M \bar{\lambda}_{ik}(x, v(x, i)) (V(x, k) - V(x, i)) + \bar{c}(x, i, v(x, i)) \\ & = \inf_{u \in U} \left\{ \sum_{j=1}^N \bar{b}_j(x, i, u) \frac{\partial V(x, i)}{\partial x_j} + \sum_{k=1}^M \bar{\lambda}_{ik}(x, u) (V(x, k) - V(x, i)) \right. \\ & \quad \left. + \bar{c}(x, i, u) \right\}, \quad \text{a.e. } x \in \mathbb{R}^N, i \in \mathcal{S}. \end{aligned}$$

Proof. The “necessity” part is contained in the proof of Theorem 6.1. We establish the sufficiency. Let $v(\cdot, \cdot)$ satisfy (6.4). The existence of such a v is guaranteed by a standard measurable selection theorem [4, Lemma 1]. Let v' be any other homogeneous Markov nonrandomized policy. By standard arguments involving Ito's formula and the strong Markov property, we conclude that

$$J_v(x, i) \leq J_{v'}(x, i), \quad \text{a.e. } x \in \mathbb{R}^N, i \in \mathcal{S}.$$

Hence, by Lemma 4.2,

$$J_v(x, i) \leq J_{\bar{v}}(x, i)$$

for any admissible policy \bar{v} . Thus, v is optimal. \square

Remark 6.1. Thus far, we have assumed that the cost function \bar{c} is bounded. However, this condition can be relaxed, as we show in the Appendix.

7. An application to a simplified model. We consider a modified version of the model studied in [2]. Suppose there is one machine producing a single commodity. We assume that the demand rate is a constant $d > 0$. Let the machine state $S(t)$ take values in $\{0, 1\}$, $S(t) = 0$ or 1 , according to whether the machine is down or functional. Let $S(t)$ be a continuous time Markov chain with generator

$$\begin{bmatrix} -\lambda_0 & \lambda_0 \\ \lambda_1 & -\lambda_1 \end{bmatrix}.$$

The inventory $X(t)$ is governed by the Ito equation

$$(7.1) \quad dX(t) = (u(t) - d) dt + \sigma dW(t),$$

where $\sigma > 0$. The production rate $u(t)$ is constrained by

$$u(t) = \begin{cases} 0 & \text{if } S(t) = 0, \\ \in [0, R] & \text{if } S(t) = 1. \end{cases}$$

Let $c : \mathbb{R} \rightarrow \mathbb{R}_+$ be the cost function which is assumed to be convex and Lipschitz continuous. Let $\alpha > 0$ be the discount factor and let $V(x, i)$ denote the value function. In this case $V(x, i)$ is the minimal nonnegative C^2 solution of the HJB equation

$$(7.2) \quad \begin{aligned} & \left(\frac{\sigma^2}{2} V''(x, 0) - dV'(x, 0) \right. \\ & \left. \frac{\sigma^2}{2} V''(x, 1) - \min_{u \in [0, R]} \{ (u - d)V'(x, 1) \} \right) \\ & = \begin{bmatrix} \lambda_0 + \alpha & -\lambda_0 \\ \lambda_1 & \alpha - \lambda_1 \end{bmatrix} \begin{pmatrix} V(x, 0) \\ V(x, 1) \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} c(x). \end{aligned}$$

Using the convexity of $c(\cdot)$, it can be shown as in [2] that $V(\cdot, i)$ is convex for each i . Hence, there exists an x^* such that

$$(7.3) \quad \begin{aligned} V'(x, 1) &\leq 0 \quad \text{for } x \leq x^*, \\ &\geq 0 \quad \text{for } x \geq x^*. \end{aligned}$$

From (7.2), it follows that the value of u which minimizes $(u - d)V'(x, 1)$ is

$$u = \begin{cases} R & \text{if } x \leq x^*, \\ 0 & \text{if } x \geq x^*. \end{cases}$$

At $x = x^*$, $V'(x^*, 1) = 0$ and therefore any $u \in [0, R]$ minimizes $(u - d)V'(x, 1)$. Thus, in view of Theorem 6.2, we can choose any $u \in [0, R]$ at $x = x^*$. To be specific, we choose $u = d$ at $x = x^*$. It follows that the following homogeneous Markov nonrandomized policy is optimal:

$$(7.4) \quad v(x, 0) = 0, \quad v(x, 1) = \begin{cases} R & \text{if } x < x^*, \\ d & \text{if } x = x^*, \\ 0 & \text{if } x > x^*. \end{cases}$$

We note at this point that the piecewise deterministic model, in general, would lead to a singular control problem when $V'(x, 1) = 0$ [2], [27]. In [2], Akella and Kumar have obtained the solution of the HJB equation (this would be (7.2) without the second-order term) in closed form and have computed an explicit expression for x^* . They have shown that a policy of the type (7.4) is optimal among all homogeneous Markov nonrandomized policies. In our case, the additive noise in (7.1) induces a smoothing effect to remove the singular situation; in addition, our results imply that the policy (7.4) is optimal among *all admissible policies*. The only limitation of our model is that it would, in general, be very difficult to solve (7.2) analytically. Therefore, we must rely on numerical methods to compute an optimal policy of the type (7.4).

We now discuss the manufacturing model studied in [27] as described in the introduction. The machine state $S(t)$ is again a prescribed continuous time Markov chain taking values in $S = \{1, \dots, M\}$. For each $i \in \mathcal{S}$, the production rate $u = (u_1, \dots, u_N)$ takes values in U_i , a convex polyhedron in \mathbb{R}^N . The demand rate is $d = [d_1, \dots, d_N]^T$. In this case, if the cost function $c : \mathbb{R}^N \rightarrow \mathbb{R}_+$ is Lipschitz continuous and convex, it can be shown that for each $i \in \mathcal{S}$, the value function $V(\cdot, i)$ is convex. But from this fact alone optimal policies of the type (7.4) cannot be obtained. However, since an optimal homogeneous Markov nonrandomized policy $v(x, i)$ is determined by minimizing

$$\sum_{j=1}^N (u_j - d_j) \frac{\partial V(x, i)}{\partial x_j}$$

over U_i , $v(x, i)$ takes values at extreme points of U_i . Thus, for each machine state i , an optimal policy divides the buffer state space into a set of regions in which the production rate is constant. If the gradient $\nabla V(x, i)$ is zero or orthogonal to a face of U_i , a unique minimizing value does not exist. But again, in view of Theorem 6.2, we may prescribe arbitrary production rates at those points where $\nabla V(x, i) = 0$, and if $\nabla V(x, i)$ is orthogonal to a face of U_i , we can choose any corner of that face. Hence, once again, we can circumvent the singular situation.

8. Concluding remarks. We have analyzed the optimal control of switching diffusions with a discounted criterion on the infinite horizon. The model allows a very general form of coupling between the continuous and the discrete components of the process. We have shown that there exists a homogeneous, nonrandomized Markov policy that is optimal in the class of all admissible policies. Also, the existence of a unique solution in a certain class to the associated HJB equations is established and the optimal policy is characterized as a minimizing selector of an appropriate Hamiltonian.

The primary motivation for this study is a class of control problems encountered in flexible manufacturing systems. By explicitly taking into account the noise present in the dynamics, we are able to remove singularities arising in the noiseless situation. In addition, we show that hedging type policies are optimal in a much wider class of nonanticipative policies than previously considered. We have confined our attention to the flow control level only. However, our results can be used to study control problems at other levels in hierarchical manufacturing systems [21], as well as control problems in other hybrid systems (see, e.g., [17], [38], [39]).

Here we have studied only the discounted criterion. Following [12], we can obtain similar results for the finite horizon and exit time criteria. However, the long-run average cost problem is more involved and is currently under study.

Appendix. Note that by the arguments of §5 we can establish the existence of a homogeneous Markov nonrandomized policy for each fixed initial law. The independence of the optimal policy of the initial law results from the dynamic programming characterization of the optimal policy via Theorem 6.2. Using probabilistic arguments, dynamic programming equations can be derived by suitably adapting the approach in [11, Chap. 3]. However, in a brief sketch we will present an alternative analytical approach, which parallels that used for classical diffusions in [29].

We assume that for each $i \in \mathcal{S}$, $\bar{c}(\cdot, i, \cdot)$ is Lipschitz in its first argument uniformly with respect to the third. We further assume that for each $x \in \mathbb{R}^N$ and $i \in \mathcal{S}$, the value function $V(x, i) < \infty$. (This assumption may be replaced by some ergodicity

hypotheses of the process under some homogeneous Markov policy.) Let $B_R = \{x \in \mathbb{R}^N : \|x\| < R\}$. Consider the Dirichlet problem on B_R

$$(A.1) \quad \begin{aligned} \inf_{u \in U} L^u \varphi(x, i) &= \alpha \varphi(x, i), \quad \text{in } B_R \times \mathcal{S}, \\ \varphi(x, i) \Big|_{\partial B_R} &= 0. \end{aligned}$$

The existence of a unique solution $\varphi_R(x, i)$ of (A.1) in $W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S})$, $2 \leq p < \infty$, is guaranteed by [31, Thm. 5.1, p. 422]. Thus, to each $R > 0$ there corresponds a solution φ_R to (A.1) belonging to $W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S})$, for $2 \leq p < \infty$. Using elliptic regularity results as in Corollary 6.1, it follows that $\varphi_R(x, i) \in C^{2,\gamma}(B_R \times \mathcal{S})$, $0 < \gamma < 1$, γ arbitrarily close to 1. Let v_R be a homogeneous Markov nonrandomized policy that is a minimizing selector in (6.4). Standard arguments involving Ito's formula yield

$$(A.2) \quad \begin{aligned} \varphi_R(x, i) &= E \left[\int_0^{\tau_R} e^{-\alpha t} \bar{c}(X(t), S(t), v_R(X(t), S(t))) dt \mid X(0) = x, S(0) = i \right] \\ &= \inf_{u(\cdot)} E \left[\int_0^{\tau_R} e^{-\alpha t} \bar{c}(X(t), S(t), u(t)) dt \mid X(0) = x, S(0) = i \right], \end{aligned}$$

where τ_R is the hitting time of ∂B_R of the process $X(\cdot)$. Clearly, $\varphi_R(x, i) \leq V(x, i)$ and it can be easily seen from (A.2) that $\varphi_R(x, i)$ is increasing in R . Let $R' > R$. Then, by the interior estimates [31, pp. 398–402], $\{\varphi_{R'}\}_{R' > R}$ is bounded in B_R uniformly in R' and $\{\nabla \varphi_{R'}\}_{R' > R}$ is bounded in $W^{1,2}(B_R \times \mathcal{S})$ uniformly in R' . By Sobolev's imbedding theorem, $W^{1,2}(B_R \times \mathcal{S}) \hookrightarrow L^{2+\varepsilon}(B_R \times \mathcal{S})$, for some $\varepsilon > 0$. Then, by suitably modifying (4.10) of [31, p. 400], we obtain

$$\|\varphi_{R'}\|_{W^{2,2+\varepsilon}(B_R \times \mathcal{S})} \leq K_R,$$

where K_R is a constant that does not depend on R' . (The modification is needed because of the factor $\varepsilon > 0$, but it is routine.) Repeating the above procedure over and over again, we conclude that $\{\varphi_{R'}\}_{R' > R}$ is uniformly bounded in $W^{2,p}(B_R)$, for $2 \leq p < \infty$. Since $W^{2,p}(B_R) \hookrightarrow W^{1,p}(B_R)$ and the injection is compact, it follows that $\{\varphi_R\}$ converges strongly in $W^{1,p}(B_R)$. Thus, given any sequence $\{R_n\}$, $R_n \rightarrow \infty$, as $n \rightarrow \infty$ and for any fixed integer $N \geq 2$, we can choose a subsequence $\{R_{n_i}\}$ such that $\{\varphi_{R_{n_i}}\}$ converges strongly in $W^{1,p}(B_{N-1})$. Using a suitable diagonalization, we may assume that $\{\varphi_{R_{n_i}}\}$ converges strongly in $W^{1,p}(B_{N-1})$ for each integer $N \geq 2$. Let ψ be a limit point of $\{\varphi_{R_{n_i}}\}$. It can be shown as in [5, p. 148] (see also [31, p. 420]) that

$$\begin{aligned} \inf_{u \in U} \left\{ \sum_{k=1}^N \bar{b}_k(x, j, u) \frac{\partial \varphi_{R_{n_i}}(x, j)}{\partial x_k} + \sum_{\ell=1}^M \bar{\lambda}_{j\ell}(x, u) (\varphi_{R_{n_i}}(x, \ell) - \varphi_{R_{n_i}}(x, j)) + \bar{c}(x, j, u) \right\} \\ \longrightarrow \inf_{u \in U} \left\{ \sum_{k=1}^N \bar{b}_k(x, j, u) \frac{\partial \psi(x, j)}{\partial x_k} + \sum_{\ell=1}^M \bar{\lambda}_{j\ell}(x, u) (\psi(x, \ell) - \psi(x, j)) + \bar{c}(x, j, u) \right\} \end{aligned}$$

strongly in $L^p(B_{N-1})$. Therefore, $\psi \in W_{loc}^{1,p}(\mathbb{R}^N \times \mathcal{S})$ and ψ satisfies

$$\inf_{u \in U} L^u \psi(x, i) = \alpha \psi(x, i)$$

in $\mathcal{D}'(\mathbb{R}^N \times \mathcal{S})$, i.e., in the sense of distributions. By elliptic regularity, $\psi \in W_{loc}^{2,p}(\mathbb{R}^N \times \mathcal{S})$, $2 \leq p < \infty$. Therefore, as in Corollary 6.1, it follows that $\psi \in C^{2,\gamma}(\mathbb{R}^N \times \mathcal{S})$, $0 < \gamma < 1$, γ arbitrarily close to 1. Let v be a minimizing selector corresponding to ψ . Then, by standard arguments involving Ito's formula, it can be shown that

$$\begin{aligned}\psi(x, i) &= E \left[\int_0^\infty e^{-\alpha t} \bar{c}(X(t), S(t), v(X(t), S(t))) dt \mid X(0) = x, S(0) = i \right], \\ &= \inf_{u(\cdot)} E \left[\int_0^\infty e^{-\alpha t} \bar{c}(X(t), S(t), u(t)) dt \mid X(0) = x, S(0) = i \right].\end{aligned}$$

Thus, $\psi(x, i) = V(x, i)$. In this situation, (6.1) does not have a unique solution in general, but $V(x, i)$ can be identified as a minimal nonnegative solution of (6.1) in $C^2(\mathbb{R}^N \times \mathcal{S})$. The assertion of Theorem 6.2 is also valid in this case.

REFERENCES

- [1] R. AKELLA, Y. CHOONG, AND S. B. GERSHWIN, *Performance of hierarchical production scheduling policy*, IEEE Trans. on Components, Hybrids and Manufacturing Technology, CHMT-7 (1984), pp. 225–240.
- [2] R. AKELLA AND P. R. KUMAR, *Optimal control of production rate in a failure prone manufacturing system*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 116–126.
- [3] D. G. ARONSON, *Bounds for the fundamental solution of a parabolic equation*, Bull. Amer. Math. Soc., 73 (1967), pp. 890–896.
- [4] V. E. BENEŠ, *Existence of optimal strategies based on specified information for a class of stochastic decision problems*, SIAM J. Control Optim., 8 (1970), pp. 179–188.
- [5] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.
- [6] A. BENSOUSSAN AND J. L. LIONS, *Impulse Control and Quasi-Variational Inequalities*, Gauthier-Villars, Paris, 1984.
- [7] ———, *Optimal control of random evolutions*, Stochastics, 5 (1981), pp. 169–199.
- [8] A. BENSOUSSAN, S. P. SETHI, R. VICKSON, AND N. DERZKO, *Stochastic production planning with production constraints*, SIAM J. Control Optim., 22 (1984), pp. 920–935.
- [9] T. BIELECKI AND P. R. KUMAR, *Optimality of zero-inventory policies for unreliable manufacturing systems*, Oper. Res., 36 (1988), pp. 532–546.
- [10] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley, New York, 1968.
- [11] V. S. BORKAR, *Optimal Control of Diffusion Processes*, Pitman Res. Notes Math. Ser., No. 203, Longman, Harlow, UK, 1989.
- [12] V. S. BORKAR AND M. K. GHOSH, *Controlled diffusions with constraints*, J. Math. Anal. Appl., 152 (1990), pp. 88–108.
- [13] R. W. BROCKETT AND G. L. BLANKENSHIP, *A representation theorem for linear differential equations with Markovian coefficients*, Proc. of 1977 Allerton Conference on Circuits and Systems Theory, Urbana, Illinois, 1977.
- [14] E. K. BOUKAS AND A. HAURIE, *Optimality conditions for continuous time systems with controlled jump Markov disturbances: application to an FMS planning problem*, Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Inform. Sci., Vol. III, Springer-Verlag, New York, 1988, pp. 633–676.
- [15] ———, *Manufacturing flow control and preventive maintenance: a stochastic control approach*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 1024–1031.
- [16] E. K. BOUKAS, A. HAURIE, AND P. MICHEL, *An optimal control problem with a random stopping time*, J. Optim. Theory Appl., 64 (1990), pp. 471–480.
- [17] D. A. CASTANON, M. CODERCH, B. C. LEVY, AND A. S. WILLSKY, *Asymptotic analysis, approximation, and aggregation methods for stochastic hybrid systems*, Proc. 1980 Joint Automatic Control Conference, San Francisco, CA, 1980.

- [18] J. CHABROWSKI AND N. A. WATSON, *Properties of solutions of weakly coupled parabolic systems*, J. London Math. Soc., 23 (1981), pp. 475–495.
- [19] M. H. A. DAVIS, *Stochastic Control and Nonlinear Filtering*, Tata Institute of Fundamental Research, Bombay, 1984.
- [20] W. H. FLEMING, S. P. SETHI, AND H. M. SONER, *An optimal stochastic production planning with randomly fluctuating demand*, SIAM J. Control Optim., 25 (1987), pp. 1494–1502.
- [21] S. B. GERSHWIN, *Hierarchical flow control: a framework for scheduling and planning discrete events in manufacturing systems*, Proc. IEEE, 77 (1989), pp. 195–209.
- [22] I. I. GIHMAN AND A. V. SKOROHOD, *Controlled Stochastic Processes*, Springer-Verlag, New York, 1979.
- [23] P. GRISVARD, *Elliptic Problems in Non-Smooth Domains*, Pitman, Boston, 1965.
- [24] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.
- [25] J. JACOD AND A. N. SHIRYAYEV, *Limit Theorems for Stochastic Processes*, Springer-Verlag, New York, 1980.
- [26] S. KESAVAN, *Topics in Functional Analysis and Applications*, Wiley, New Delhi, 1989.
- [27] J. KIMENIA AND S. B. GERSHWIN, *An algorithm for the computer control of a flexible manufacturing system*, Institute of Industrial Engineers Trans., 15 (1983), pp. 353–362.
- [28] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
- [29] H. J. KUSHNER, *Optimal discounted stochastic control for diffusion processes*, SIAM J. Control, 5 (1967), pp. 520–531.
- [30] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc., Providence, RI, 1968.
- [31] O. A. LADYZHENSKAYA AND N. N. URAL'CEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
- [32] J. LEHOCZKY, S. SETHI, H. M. SONER, AND M. TAKSAR, *An asymptotic analysis of hierarchical control of manufacturing systems under uncertainty*, Math. Oper. Res., 16 (1991), pp. 596–608.
- [33] G. J. OLDSER AND R. SURI, *Time optimal control of parts-routing in a manufacturing system with failure prone machines*, Proc. 19th IEEE Conference on Decision and Control, Albuquerque, NM, 1980, pp. 722–727.
- [34] R. PHELPS, *Lectures on Choquet's Theorem*, Van Nostrand, New York, 1966.
- [35] R. RISHEL, *Dynamic programming and minimum principles for systems with jump Markov disturbances*, SIAM J. Control Optim., 13 (1975), pp. 338–371.
- [36] S. SETHI AND M. I. TASKAR, *Deterministic equivalent for a continuous time linear-convex stochastic control problem*, J. Optim. Theory Appl., 64 (1990), pp. 169–181.
- [37] A. SHARIFNIA, *Production control of a manufacturing system with multiple machine states*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 620–625.
- [38] D. D. SWORDER, *Feedback control of a class of linear systems with jump parameters*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 9–14.
- [39] ———, *Control of systems subject to sudden change in character*, Proc. IEEE, 64 (1976), pp. 1219–1225.
- [40] ———, *Control of a linear system with non-Markovian modal changes*, J. Econom. Dynamics Control, 2 (1980), pp. 233–240.
- [41] A. JU. VERETENNIKOV, *On strong solutions of Ito stochastic equations with jumps*, Theory Probab. Appl., 32 (1988), pp. 148–152.
- [42] D. VERMES, *Optimal control of piecewise deterministic Markov processes*, Stochastics, 14 (1985), pp. 165–208.
- [43] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, A. T. Bharucha-Reid, ed., Vol. 2, Academic Press, New York, 1970, pp. 131–212.

ON DIFFERENTIAL SYSTEMS WITH QUADRATIC IMPULSES AND THEIR APPLICATIONS TO LAGRANGIAN MECHANICS*

ALBERTO BRESSAN[†] AND FRANCO RAMPAZZO[‡]

Abstract. This paper is concerned with the basic dynamics and a class of variational problems for control systems of the form

$$(E) \quad \dot{x} = f(t, x, u) + g(t, x, u)\dot{u} + h(t, x, u)\dot{u}^2$$

These systems have impulsive character, due to the presence of the time derivative \dot{u} of the control. It is shown that trajectories can be well defined when the controls u are limits (in a suitable weak sense) of sequences (u_n) contained in the Sobolev space $W^{1,2}$. Roughly speaking, one can say that, in this case, the \dot{u}_n tend to the square root of a measure. Actually, this paper shows that the system (E) is essentially equivalent to an (affine) impulsive system of the form

$$\dot{x} = f(x) + g(x)v + h(x)\dot{w},$$

where $v \in L^2$ and \dot{w} is a nonnegative Radon measure not smaller than v^2 . This provides a characterization of the closure of the set of trajectories of (E), as the controls u range inside a fixed ball of $W^{1,2}$. The existence of (generalized) optimal controls for variational problems of Mayer type is also investigated. Since the main motivation for studying systems of form (E) comes from Rational Mechanics, this paper concludes by presenting an example of an impulsive Lagrangian system.

Key words. nonlinear impulsive system, L^1 convergence of solutions, existence of optimal generalized controls, Lagrangian mechanical system

AMS subject classifications. 34K35, 49A10, 70D10

1. Introduction and notation. To a mechanical Lagrangian system with $N+1$ degrees of freedom and locally parametrized by coordinates (q_i, u) , $i = 1, \dots, N$, one can superimpose new time-dependent constraints by assigning the motion $u(\cdot)$ of the coordinate u . This leads to *impulsive control systems* of the form

$$(E) \quad \dot{x} = f(t, x, u) + g(t, x, u)\dot{u} + h(t, x, u)\dot{u}^2, \quad x(0) = \bar{x} \in R^n \quad t \in [0, T],$$

where, as customary, the dot denotes differentiation with respect to time. The appellation *impulsive control systems* is here adopted because of the presence of the derivative of the control on the right-hand sides of the equations. We refer to §6 and to the references quoted there for a detailed description of the mechanical problems that motivate the study of equations having the form (E). In particular, (6.1) provides a concrete example where the derivative of the control appears quadratically. This equation concerns a swing where, unlike the situation considered in [9], the role of control is played by the angle formed by the swing and the vertical.

We remark that in the usual applications of control theory to classical mechanics the quantities identified with the controls have the character of forces or torques. This leads to ordinary control equations, which do not contain the derivative of the

*Received by the editors February 6, 1991; accepted for publication (in revised form) April 21, 1991.

[†]Scuola Internazionale Superiore di Studi Avanzati, Via Beirut 2, 34014 Trieste, Italy.

[‡]Department of Mathematics, University of Padova, Via Belzoni 7, 35131 Padova, Italy.

control. On the contrary, here the control is one of the positional coordinates of the mechanical system Σ . For example, one could try to *guide* the entire motion of Σ just by assigning the motion of a part Π of Σ : in this case what one can practically measure is the time evolution of the positions of Π , while one does not know the forces necessary to generate such evolution, as it would be required by the classical approach of control theory to mechanics.

The important case $h \equiv 0$, which occurs in such mechanical applications under special assumptions on the geometry of the constraint manifold [14], [15], was considered in [2]–[5], [12], [16]–[18].

The aim of this paper is to provide a definition of solution of (E) valid for the general case, i.e., when the derivative \dot{u} appears quadratically, and for a suitably large class of inputs u . We study the existence, uniqueness, and continuous dependence (on the controls) for the corresponding Cauchy problem. This allows us to drop the aforementioned technical assumptions on the geometry of the constraint manifold and to cover a very large class of mechanical applications. In particular, all the problems where the forces depend on the velocities linearly and quadratically are within this class, without any assumption on the geometry of the constraints.

Observe first, that by possibly introducing the variables $x^0 = t$, $x^{n+1} = u$ and the new equations $\dot{x}^0 = 1$, $\dot{x}^{n+1} = \dot{u}$, it is not restrictive to assume that (E) takes the simpler form

$$(1.1) \quad \dot{x} = f(x) + g(x)\dot{u} + h(x)\dot{u}^2, \quad x(0) = \bar{x} \in R^n.$$

When u is absolutely continuous with square integrable derivative, the solutions of (1.1) are already well defined. In the special case, where $h \equiv 0$, it is known that the input-output map $\psi: u(\cdot) \rightarrow x(\cdot)$ can be extended continuously to a map from $L^1([0, T], R)$ into $L^1([0, T], R^n)$. However, as shown in [13], when $h \neq 0$, such continuous extension does not exist. On the other hand, as u ranges within bounded subsets of the Sobolev space $W^{1,2}([0, T], R)$, we show in §2 that the input-output map remains continuous also with respect to a new topology \mathcal{T} on the space of controls, somewhat weaker than the norm topology. This enables us to embed the map ψ in a larger domain consisting of couples (v, \dot{w}) , where $v \in L^2([0, T], R)$ and w is a nondecreasing function with the property

$$\int_a^b v^2(t) \, dt \leq w(b) - w(a) \quad \text{for all } 0 \leq a < b \leq T.$$

For any such couple we consider a problem of the form

$$(1.2) \quad \dot{x} = f(x) + g(x)v + h(x)\dot{w}, \quad x(0) = \bar{x}.$$

We prove that the solution of (1.2) depends continuously on the input pair (v, w) , where $v \in L^2_{\text{weak}}([0, T], R)$ and $w \in L^1([0, T], R)$. This also provides a characterization of the L^1 -closure of the set of trajectories of (1.1) whenever u ranges inside a fixed ball in $W^{1,2}([0, T], R)$.

Furthermore, we prove that the set reached at a time $t = \tau$, $\tau \in [0, T]$, by (generalized) solutions of (1.2) is compact. This guarantees the existence of (generalized) optimal controls for a class of variational problems having a dynamics of the form (E).

In §4, some examples display the difficulties that arise when no bound on the L^2 norm of \dot{u} is imposed. Still, if the vector field h satisfies a suitable coercivity

condition, then $|x(t)| \rightarrow \infty$ as $\int_0^t |\dot{u}(s)|^2 ds \rightarrow \infty$; hence, the boundedness of solutions $x(\cdot)$ automatically implies a bound on $\|\dot{u}\|_{L^2}$.

Finally, an application concerning an optimal control problem for a mechanical system is presented in §5.

The following is a list of the notation that will be used in the next sections.

The euclidean norm of a vector $y \in R^n$ is written $|y|$, while $B[u, \rho]$ denotes the closed ball centered at y with radius ρ . For $r = 1, 2, \dots$ [$r = 0$], we say that a map $f : R^\ell \mapsto R^m$ is of class C^r if f is r times continuously differentiable [if f is continuous]. When f is at least C^1 , by $Df(x)$ we denote the $m \times \ell$ Jacobian matrix of f at the point $x \in R^\ell$. $|Df(x)|$ indicates the operator norm of $Df(x)$.

Let h be a C^1 vector field on R^d . As customary, $(\exp th)(x)$ indicates the value at time t of the solution of the Cauchy problem

$$\dot{y} = h(y), \quad y(0) = x.$$

The differential of the map $x \mapsto (\exp th)(x)$ is denoted by $(\exp th)_*$. Observe that, for each $x \in R^d$, $(\exp th)_*$ is a linear map from the tangent space at x into the tangent space at the point $(\exp th)(x)$. In particular, if f is a vector field on R^d , one has

$$(1.3) \quad (\exp th)_* f(x) = \lim_{\varepsilon \rightarrow 0} \frac{(\exp th)(x + \varepsilon f(x)) - (\exp th)(x)}{\varepsilon} = v(t),$$

where $v(\cdot)$ is the solution of the linear Cauchy problem

$$(1.4) \quad \dot{v}(s) = Dh((\exp sh)(x)) \cdot v(s), \quad v(0) = f(x).$$

In the following, $W^{r,p}([a, b], R^d)$ is the Sobolev space of functions from $[a, b]$ into R^d whose distributional derivatives up to order r belong to L^p . If w is a map from $[a, b]$ into R^d , the total variation of w on the interval $[a, b]$ is written $V_{[a,b]}(w)$. We recall that

$$V_{[a,b]}(w) = \sup \left(\sum_{i=1}^n |f(t_i) - f(t_{i-1})| \right),$$

where the supremum is taken over all partitions $\{a = t_0 < t_1 < \dots < t_n = b\}$ of the interval $[a, b]$.

If E is a topological vector space, by E^* we denote its topological dual space. The convergence of a sequence $(v_n)_{n \geq 1}$ to a point v in E is written as $v_n \rightarrow v$, while “ \rightharpoonup ” indicates weak convergence.

2. Systems with linear impulses. As a preliminary, we provide a notion of generalized solution to the impulsive Cauchy problem

$$(2.1) \quad \begin{cases} \dot{x} = f(x) + g(x)v + h(x)\dot{w} \\ x(0) = \bar{x}, \end{cases}$$

corresponding to a control pair $(v, w)(\cdot)$, in the case where $v \in L^1([0, T], R)$ and w is a bounded, measurable function. Concerning the vector fields f, g, h we shall assume

(H) The maps f, g are C^1 . The vector field h is C^2 and *complete*, i.e., for every x the map $t \rightarrow (\exp th)(x)$ is defined for all $t \in R$.

Recalling the notation at (1.3), for a given $w : [0, T] \mapsto R$ consider the vector fields

$$f^w(t, y) \doteq (\exp(-w(t)h))_* f((\exp w(t)h)(y)),$$

$$g^w(t, y) \doteq (\exp(-w(t)h))_* g((\exp w(t)h)(y)).$$

DEFINITION 2.1. A map $x : [0, T] \rightarrow R^n$ is a *generalized solution* of (2.1) if

$$(2.2) \quad x(t) = (\exp w(t)h)(y(t)) \quad \forall t \in [0, T],$$

where $y(\cdot)$ is a Caratheodory solution of the Cauchy problem

$$(2.3) \quad \begin{cases} \dot{y}(t) = f^w(t, y(t)) + g^w(t, y(t)) \cdot v(t), \\ y(0) = (\exp(-w(0)h))(\bar{x}). \end{cases}$$

Remark 2.1. If w is smooth, the above definition is equivalent to the classical one. Theorem 2.2 will show that this is the unique extension to the case where $v \in L^1$ and w is bounded measurable.

Remark 2.2. If $w' : [0, T] \mapsto R$ is a measurable function such that $w'(0) = w(0)$ and $w'(t) = w(t)$ almost everywhere, then $f^{w'}(t, \cdot) = f^w(t, \cdot)$, $g^{w'}(t, \cdot) = g^w(t, \cdot)$ for almost every t . Hence, the corresponding solutions y' , y of (2.3) coincide. The value of the solution $x(\cdot)$ of (2.1) at a given time τ thus depends only on the L^1 equivalence class of v, w and on the values of w at $t = 0$ and at $t = \tau$.

In the following, when we are interested only in the value of a solution at a single time τ , we thus consider functions v, w defined up to L^1 equivalence on $[0, T]$, with w pointwise determined at $t = 0, \tau$.

THEOREM 2.1. *If the hypotheses (H) hold, then there exists $\tau > 0$ such that the Cauchy problem (2.1) has a unique generalized solution on $[0, \tau]$. If, in addition, the vector fields f, g, h satisfy the bounds*

$$(2.4) \quad |f(x)| \leq C(1 + |x|), \quad |g(x)| \leq C(1 + |x|), \quad |Dh(x)| \leq C$$

for some constant C and all $x \in R^n$, then the generalized solution of (2.1) exists on the whole interval $[0, T]$ and is uniquely defined.

Proof. By definition, it is clear that the generalized solution $x(\cdot)$ of (2.1) exists and is unique if and only if the same holds for the solution $y(\cdot)$ of (2.3). For any measurable bounded $w(\cdot)$, the functions f^w, g^w satisfy the hypotheses of Caratheodory's theorem. being measurable with respect to t and continuously differentiable with respect to y . Therefore, a unique local solution of (2.3) exists.

Concerning the global existence, if M_0, M_1 are constants for which

$$(2.5) \quad |h(0)| = M_0, \quad |w(t)| \leq M_1 \quad \forall t \in [0, T],$$

then the bounds (2.4), (2.5) imply

$$|h(y)| \leq M_0 + C|y|, \quad |(\exp(-w(t)h))_*| \leq e^{C|w(t)|},$$

$$|(\exp w(t)h)(y)| \leq e^{CM_1}|y| + \frac{M_0}{C}(e^{CM_1} - 1),$$

$$(2.6) \quad \begin{aligned} |f^w(t, y)| &= |(\exp(-w(t)h))_* f((\exp w(t)h)(y))| \\ &\leq e^{CM_1} \cdot C \left(1 + e^{CM_1}|y| + \frac{M_0}{C}(e^{CM_1} - 1) \right) \leq C_1(1 + |y|) \end{aligned}$$

for some constant C_1 . Similarly,

$$(2.7) \quad |g^w(t, y)| \leq C_2(1 + |y|).$$

Recalling that $v \in L^1$, the bounds (2.6) and (2.7) imply the global existence of the solution $y(\cdot)$ of (2.3), on the interval $[0, T]$.

The following approximation result justifies the notion of generalized solution introduced in Definition 2.1.

THEOREM 2.2. *Assume that the hypotheses (H) and (2.4) on the vector fields f, g, h hold, and let (v, w) be a control pair in $L^2([0, T], R) \times L^\infty([0, T], R)$, with w pointwise defined at a given $\tau \in [0, T]$ and at $t = 0$. Let $(v_n, w_n)_{n \in \mathbf{N}}$ be a sequence of control pairs in $L^2 \times W^{1,2}$, such that $\|v_n\|_{L^2} \leq L$, $V_{[0,T]}(w_n) \leq L$, for some constant L , and*

$$\begin{aligned} v_n &\rightharpoonup v \quad \text{in } L^2([0, T], R), \\ w_n &\rightarrow w \quad \text{in } L^1([0, T], R). \end{aligned}$$

Moreover, suppose that

$$w_n(0) \rightarrow w(0), \quad w_n(\tau) \rightarrow w(\tau).$$

Then the (Caratheodory) solutions x_n of (2.1) corresponding to the control pairs (v_n, w_n) satisfy

$$(2.8) \quad \begin{aligned} x_n &\rightarrow x \quad \text{in } L^1([0, T], R^n), \\ x_n(\tau) &\rightarrow x(\tau), \end{aligned}$$

where $x(\cdot)$ denotes the generalized solution of (2.1) corresponding to (v, w) , pointwise defined at $t = \tau$ and $t = 0$.

Proof. In order to prove (2.8) it will be shown that from every subsequence $(x_{n'})_{n' \geq 1}$ of $(x_n)_{n \geq 1}$ one can extract a further subsequence $(x_\nu)_{\nu \geq 1}$ converging to x in $L^1([0, T], R^n)$ and pointwise at τ .

For every $n' \geq 1$, let $y_{n'}$ be the solution of (2.6) corresponding to the control pair $(v_{n'}, w_{n'})$. By the hypotheses on f, g, h and on v_n, w_n , the variations $V_{[0,T]}(y_{n'})$ are uniformly bounded. Hence, by Helley's theorem, there exists a subsequence $(y_\nu)_{\nu \geq 1}$ converging to a map y_∞ pointwise on $[0, T]$. By possibly extracting a further subsequence, we can also assume $w_\nu(t) \rightarrow w(t)$ for almost every t . We claim that y_∞ coincides with the solution y corresponding to the control pair (v, w) . By the uniqueness of solutions of (2.3), this can be established by showing that

$$y_\infty(t) = y_\infty(0) + \int_0^t [f^w(s, y_\infty(s)) + g^w(s, y_\infty(s))v(s)] \, ds$$

for every $t \in [0, T]$. In fact, we have

$$(2.9) \quad \begin{aligned} &\left| y_\infty(t) - y_\infty(0) - \int_0^t [f^w(s, y_\infty(s)) + g^w(s, y_\infty(s))v(s)] \, ds \right| \\ &\leq |y_\infty(t) - y_\nu(t)| + |y_\infty(0) - y_\nu(0)| + \int_0^t |f^w(s, y_\infty(s)) - f^{w_\nu}(s, y_\nu(s))| \, ds \\ &\quad + \left| \int_0^t g^w(s, y_\infty(s))[v(s) - v_\nu(s)] \, ds \right| + \int_0^t |g^w(s, y_\infty(s)) - g^{w_\nu}(s, y_\nu(s))| \cdot |v_\nu(s)| \, ds. \end{aligned}$$

For every t , as $\nu \rightarrow \infty$, the first two terms on the right-hand side of (2.9) converge to zero. By construction we have $f^{w_\nu}(s, y_\nu(s)) \rightarrow f^w(s, y_\infty(s))$ for almost every s . Hence, by the Dominated Convergence theorem, the third term on the right-hand side of (2.9) also converges to zero. The fourth term converges to zero because $g^w(\cdot, y_\infty(\cdot))$ is a bounded function, and the v_ν converge to v weakly in L^2 . Finally, the Hölder inequality

$$\begin{aligned} \int_0^t |g^w(s, y_\infty(s)) - g^{w_\nu}(s, y_\nu(s))| \cdot |v_\nu(s)| \, ds \\ \leq \|g^w(\cdot, y_\infty(\cdot)) - g^{w_\nu}(\cdot, y_\nu(\cdot))\|_{L^2} \cdot \|v_\nu\|_{L^2} \end{aligned}$$

implies that also the fifth term on the right-hand side of (2.9) converges to zero. Indeed, $\|v_\nu\|_{L^2} \leq L$, and by dominated convergence, $g^{w_\nu}(\cdot, y_\nu(\cdot)) \rightarrow g^w(\cdot, y_\infty(\cdot))$ in L^2 .

From the equality $y_\infty(t) = y(t)$ for every $t \in [0, T]$, since $w_\nu(t) \rightarrow w(t)$ almost everywhere and at $t = \tau$, we obtain that

$$x_\nu(t) = (\exp w_\nu(t)h)(y_\nu(t)) \rightarrow (\exp w(t)h)(y(t)) = x(t)$$

almost everywhere and at $t = \tau$. This implies $x_\nu \rightarrow x$ in $L^1([0, T]; \mathbb{R}^n)$, because the x_ν are uniformly bounded. Since the subsequence $(x_{n'})_{n' \geq 1}$ was arbitrary, the theorem is proved.

3. Quadratic impulses. Consider again the Cauchy problem (1.1) and, for a fixed positive constant K , define the family of controls

$$U_K = \{u \in W^{1,2}([0, T], \mathbb{R}), \quad u(0) = 0, \quad \|\dot{u}\|_{L^2} \leq \sqrt{K}\}.$$

Observe that, if the vector fields f, g, h are continuously differentiable, then for every $u \in W^{1,2}$, a local Caratheodory solution to the Cauchy problem (1.1) exists. If, in addition, the assumptions (H) and (2.4) hold, then for every $u \in U_K$, the solution of (1.1) is defined on the entire interval $[0, T]$.

Denoting by X_K the set of solutions of (1.1) corresponding to the controls of U_K , we wish to characterize the L^1 -closure of X_K . For this purpose, consider the set \tilde{U}_K formed by the control pairs (v, w) such that:

- (i) The function v lies in L^2 , with $\|v\|_{L^2}^2 \leq K$.
- (ii) $w : [0, T] \rightarrow \mathbb{R}$ is nondecreasing, with $w(0) = 0$, $w(T) \leq K$.
- (iii) For every $[a, b] \subseteq [0, T]$ we have

$$(3.1) \quad \int_a^b v^2(t) \, dt \leq w(b) - w(a).$$

According to the results of the previous section, for every $(v, w) \in \tilde{U}_K$, there exists a generalized solution of (2.1). Let us denote by \tilde{X}_K the set of the generalized solutions of (2.1) corresponding to control pairs in \tilde{U}_K . The next two theorems show that \tilde{X}_K is a compact subset of $L^1([0, T], \mathbb{R}^n)$ and coincides with the closure of X_K . In the following, we assume that the vector fields f, g, h satisfy the assumptions (H) and (2.4).

THEOREM 3.1. *Let $(u_n)_{n \geq 1}$ be a sequence in U_K , and let $x(u_n)$ be the corresponding solutions of (1.1). Then, there exists a subsequence $(u_{\nu})_{\nu \geq 1}$ such that the $x(u_\nu)$ converge to some $x \in \tilde{X}_K$ in $L^1([0, T], \mathbb{R}^n)$ and pointwise everywhere on $[0, T]$.*

THEOREM 3.2. *If $x \in \tilde{X}_K$, then there exists a sequence $(u_n)_{n \geq 1}$ in U_K such that the corresponding solutions $x(u_n)$ of (1.1) converge to x in $L^1([0, T], \mathbb{R}^n)$ and pointwise everywhere on $[0, T]$.*

Proof of Theorem 3.1. By hypothesis, the maps $v_n = \dot{u}_n$ belong to the closed ball $B[0, \sqrt{K}]$ in L^2 . Since this ball is weakly compact, there exists a subsequence $(v_{n'})_{n' \geq 1}$ such that $v_{n'} \rightharpoonup v$ for some $v \in B[0, \sqrt{K}]$.

Next, consider the sequence of nondecreasing functions $w_{n'} : [0, T] \mapsto [0, K]$, defined as

$$w_{n'}(t) = \int_0^t v_{n'}^2(s) \, ds.$$

By Helley's theorem, one can extract a further subsequence, say w_ν , such that

$$w_\nu(t) \rightarrow w(t) \quad \forall t \in [0, T]$$

for some nondecreasing function $w : [0, T] \mapsto [0, K]$, with $w(0) = 0$. We claim that the control pair (v, w) lies in \tilde{U}_K . The properties (i) and (ii) are already clear. Moreover, for every $[a, b] \subseteq [0, T]$

$$(3.2) \quad w(b) - w(a) = \lim_{\nu \rightarrow \infty} \int_a^b v_\nu^2(t) \, dt \geq \int_a^b v^2(s) \, ds,$$

because the functional $v \mapsto \|v\|_{L^2}^2$ is convex, hence weakly lower semicontinuous (see [11], Prop. III.12). This proves (iii).

Now let $x \in \tilde{X}_K$ be the solution of (2.1) corresponding to the control pair $(v, w) \in \tilde{U}_K$. Observe that the solutions $x(u_\nu)$ of (1.1) satisfy

$$\begin{cases} \dot{x}_\nu = f(x_\nu) + g(x_\nu)v_\nu + h(x_\nu)\dot{w}_\nu, \\ x_\nu(0) = \bar{x}. \end{cases}$$

By Theorem 2.2, from the weak convergence $v_\nu \rightharpoonup v$, the pointwise convergence $w_\nu(t) \rightarrow w(t)$ for all $t \in [0, T]$ and the uniform boundedness of the sequence w_ν , it follows that $x_\nu(\cdot) \rightarrow x(\cdot)$ pointwise everywhere on $[0, T]$ and in $L^1([0, T]; \mathbb{R}^n)$. \square

Proof of Theorem 3.2. Define the functions

$$p(t) = \int_0^t v^2(s) \, ds, \quad \psi(t) = w(t) - p(t).$$

Observe that p, ψ are nondecreasing, with $p(0) = \psi(0) = 0$, $\psi(T) \leq w(T) \leq K$. Since ψ has bounded variation, it can have at most countably many points of discontinuity, say $\{\tau_i; i \geq 1\}$. Moreover, let $\{\xi_i; i \geq 1\}$ be any sequence of points dense on $[0, T]$. For every $n \geq 1$, consider the set

$$S_n = \{0, T\} \cup \{\tau_1, \dots, \tau_n\} \cup \{\xi_1, \dots, \xi_n\}.$$

Arrange S_n in increasing order, say

$$S_n = \{0 = t_{n,0} < t_{n,1} < \dots < t_{n,\nu_n} = T\},$$

where, of course, $\nu_n \leq 2n + 1$. For each $i \in \{1, \dots, \nu_n\}$, set

$$t_{n,i}^- = t_{n,i} - \frac{t_{n,i} - t_{n,i-1}}{n}, \quad \omega_{n,i} = \sqrt{n \cdot \frac{\psi(t_{n,i}) - \psi(t_{n,i-1})}{t_{n,i} - t_{n,i-1}}}$$

and define the functions

$$\begin{aligned} v_n(t) &= \begin{cases} 0 & \text{if } t \in [t_{n,i}^-, t_{n,i}] \text{ for some } i, \\ v(t) & \text{otherwise,} \end{cases} \\ \psi_n(t) &= \begin{cases} \omega_{n,i} & \text{if } t \in [t_{n,i}^-, t_{n,i}] \text{ for some } i, \\ 0 & \text{otherwise,} \end{cases} \\ u_n(t) &= \int_0^t [v_n(s) + \psi_n(s)] \, ds, \\ w_n(t) &= \int_0^t \dot{u}_n^2(s) \, ds. \end{aligned}$$

Observe that the above definitions imply

$$\begin{aligned} v_n(t)\psi_n(t) &= 0, \quad \dot{u}_n^2(t) = v_n^2(t) + \psi_n^2(t) \quad \forall t \in [0, T]. \\ \int_{t_{n,i-1}}^{t_{n,i}} \psi_n^2(t) \, dt &= \psi(t_{n,i}) - \psi(t_{n,i-1}) \quad \forall i \in \{1, \dots, \nu_n\}. \end{aligned}$$

By construction, the solution $x(u_n)$ then solves

$$\dot{x} = f(x) + g(x)(v_n + \psi_n) + h(x)\dot{w}_n.$$

In order to apply Theorem 2.2 and conclude that $x(u_n) \rightarrow x$ pointwise everywhere on $[0, T]$, we need to show that

$$(3.3) \quad v_n + \psi_n \rightharpoonup v \quad \text{in } L^2,$$

$$(3.4) \quad w_n(t) \rightarrow w(t) \quad \forall t \in [0, T].$$

Observing that $0 \leq v_n^2 \leq v^2 \in L^1$ and that

$$\text{meas} \left(\bigcup_{i=1}^{\nu_n} [t_{n,i}^-, t_{n,i}] \right) = \frac{T}{n},$$

by the dominated convergence theorem we conclude that $v_n \rightarrow v$ in the L^2 norm.

To show that $\psi_n \rightarrow 0$, fix any $\phi \in L^2([0, T]; \mathbb{R})$. For any $\varepsilon > 0$, choose $\phi' \in C([0, T]; \mathbb{R})$ such that $\|\phi' - \phi\|_{L^2} \leq \varepsilon$. Then

$$\begin{aligned} \limsup_{n \rightarrow \infty} \int_0^T \psi_n \phi \, dt &\leq \limsup_{n \rightarrow \infty} \int_0^T \psi_n \phi' \, dt + \limsup_{n \rightarrow \infty} \int_0^T \psi_n (\phi' - \phi) \, dt \\ &\leq \limsup_{n \rightarrow \infty} \|\psi_n\|_{L^1} \cdot \|\phi'\|_{L^\infty} + \limsup_{n \rightarrow \infty} \|\psi_n\|_{L^2} \cdot \|\phi' - \phi\|_{L^2} \leq 0 + \sqrt{K}\varepsilon. \end{aligned}$$

Since ε was arbitrary, this proves (3.3).

To prove (3.4), consider first a point t where w is discontinuous. Then $t = \tau_\nu$ for some ν . Hence, for every $n \geq \nu$ we have

$$\int_0^t \psi_n^2(s) \, ds = \psi(t),$$

while

$$\lim_{n \rightarrow \infty} \int_0^t v_n^2(s) \, ds = \int_0^t v^2(s) \, ds.$$

Hence, (3.4) holds. The same argument can of course be applied when $t = 0$ or $t = T$.

Next, consider the case where w is continuous at t . For any $\varepsilon > 0$, choose $\xi_\mu < t < \xi_\nu$ such that

$$w(t) - \varepsilon \leq w(\xi_\mu) \leq w(\xi_\nu) \leq w(t) + \varepsilon.$$

If $n \geq \max\{\mu, \nu\}$, then

$$\int_0^{\xi_\mu} \psi_n^2(s) \, ds = \int_0^{\xi_\mu} \psi(s) \, ds, \quad \int_0^{\xi_\nu} \psi_n^2(s) \, ds = \int_0^{\xi_\nu} \psi(t) \, dt.$$

Hence,

$$\begin{aligned} w(t) - \varepsilon \leq w(\xi_\mu) &= \lim_{n \rightarrow \infty} w_n(\xi_\mu) \leq \liminf_{n \rightarrow \infty} w_n(t) \\ &\leq \limsup_{n \rightarrow \infty} w_n(t) \leq \lim_{n \rightarrow \infty} w_n(\xi_\nu) = w(\xi_\nu) \leq w(t) + \varepsilon. \end{aligned}$$

Since ε was arbitrary, (3.4) is proved in this case as well.

An application of Theorem 2.2 now implies $x(u_n)(t) \rightarrow x(t)$ for every $t \in [0, T]$, and hence, $x(u_n) \rightarrow x$ in L^1 , by dominated convergence. \square

The above results motivate the following definition.

DEFINITION 3.1. For every $K > 0$, \tilde{U}_K is called the set of U_K -generalized controls for (1.1), and \tilde{X}_K is called the set of U_K -generalized trajectories of (1.1).

By Theorems 3.1–3.2, a bounded trajectory x is a U_K -generalized solution of (1.1) if and only if x is the pointwise limit of (Caratheodory) solutions of (1.1) corresponding to controls in U_K .

In connection with the system (1.1), for any $t \in [0, T]$ and $K > 0$, we also define the reachable set $R_K(t)$ at time t , with generalized controls in \tilde{U}_K :

$$R_K(t) \doteq \{x(t); \, x \text{ is a solution of (2.1), with } (v, w) \in \tilde{U}_K\}.$$

From the previous analysis it follows

COROLLARY 3.1. Let the vector fields f, g, h satisfy the assumptions (H) and (2.4). Then, for every $\tau \in [0, T]$ and $K > 0$, the reachable set $\mathcal{R}_K(\tau)$ is compact.

Indeed, let $(\hat{x}_n)_{n \geq 1}$ be a sequence of points in $R_K(t)$. By definition, there exists $(w_n, v_n) \in \tilde{U}_K$ such that $\hat{x}_n = x(v_n, w_n)(t)$. Choose a subsequence $(v_\nu, w_\nu)_{\nu \geq 1}$ such that, for some $(v, w) \in \tilde{U}_K$, one has

$$v_\nu \rightarrow v \quad \text{and} \quad w_\nu(t) \rightarrow w(t) \quad \forall t \in [0, T].$$

This is possible because the ball $B[0, \sqrt{K}]$ in L^2 is weakly compact and because of Helly's theorem. The inequality

$$\int_a^b v^2(s) \, ds \leq w(b) - w(a)$$

is then established as in (2.4). By Theorem 2.2 we now have $x(v_\nu, w_\nu)(t) \rightarrow x(v, w)(t)$ for every t . Hence, in particular, the sequence \hat{x}_ν converges to $x(v, w)(t) \in R_K(t)$, proving that $R_K(T)$ is compact.

Remark 3.1. The above results remain true if, in place of the hypothesis (2.4), one assumes that all trajectories of (1.1) with $u \in U_K$ remain inside a fixed compact set S . Indeed, one can then replace the vector fields f, g, h with

$$\tilde{f} = f \cdot \varphi, \quad \tilde{g} = g \cdot \varphi, \quad \tilde{h} = h \cdot \varphi,$$

where φ is a C^∞ scalar function with compact support such that

$$\varphi(x) \equiv 1 \quad \text{if } x \in S.$$

Then we can apply all previous results to the system

$$\dot{x} = \tilde{f}(x) + \tilde{g}(x)\dot{u} + \tilde{h}(x)\dot{u}^2.$$

Remark 3.2. In particular, Corollary 3.1 guarantees the existence of an optimal control for a variational problem of the form

$$(3.5) \quad \min \{ \Phi(x(\tau)); \quad x \in \tilde{U}_K \},$$

where $\Phi : R^n \mapsto R$ is lower semicontinuous. Indeed, on the compact set $R_K(\tau)$, the function Φ does attain its global minimum.

4. Controls with bounded variation. By Corollary 3.1, the set $\mathcal{R}(T)$ reached at $t = T$ by the U_K -generalized solutions of (1.1) defined at $t = T$ is compact. This is no longer true if instead of a bound on the L^2 norm of \dot{u} only a bound on the total variation $V_{[0,T]}(u)$, i.e., on the L^1 norm of \dot{u} , is imposed: for example, consider the control system

$$\begin{cases} \dot{x} = \dot{u}^2 \\ x(0) = 0 \end{cases}$$

on the time interval $[0, 1]$ and the sequence $(u_n)_{n \geq 1}$ of controls defined by

$$(4.1) \quad u_n(t) = \begin{cases} nt & \text{if } t \in [0, 1/n], \\ 1 & \text{if } t \in (1/n, 1]. \end{cases}$$

For every n , one trivially has $V_{[0,T]}(u_n) = \|\dot{u}_n\|_{L^1} = 1$. On the other hand, $x_n(1) = n$ diverges to $+\infty$.

It may also happen that the trajectories of (1.1) and the variations of the controls are uniformly bounded, while the L^2 norms of the derivatives of the controls diverge to $+\infty$.

For example, consider the control system

$$(4.2) \quad \begin{cases} \dot{x} = h(x)\dot{u}^2 \\ x(0) = \bar{x} \in R^2, \end{cases}$$

where $0 < |\bar{x}| < 1$ and h is the vector field on R^2 defined by

$$h(x) = \begin{pmatrix} -x_2 + x_1(1 - x_1^2 - x_2^2) \\ +x_1 + x_2(1 - x_1^2 - x_2^2) \end{pmatrix}.$$

For a positive constant C , consider the set of admissible controls

$$(4.3) \quad U_C^1 = \{u \mid u \in W^{1,2}([0, T], R), \quad u(0) = 0 \quad V_{[0, T]}(u) = \|\dot{u}\|_1 \leq C\}.$$

It is easy to check that for every control $u \in U_C^1$ the corresponding solution of (4.2) remains inside the closed ball $B[0; 1]$. As a matter of fact, the circle of equation $x_1^2 + x_2^2 = 1$ is the ω -limit of all trajectories of h starting outside the origin. On the other hand, the controls u_n defined at (4.1) belong to U_C^1 with $C = 1$, while their L^2 norms $\|\dot{u}_n\|_2 = n$ diverge to $+\infty$.

In the case where all positive orbits of h diverge to infinity, however, any bound on the solutions of (1.1) implies that the L^2 norms of the derivatives \dot{u} are uniformly bounded as well. Precisely, we have:

THEOREM 4.1. *Assume that for every $x \in B[0; R]$ one has*

$$(4.4) \quad \lim_{t \rightarrow +\infty} |(\exp th)(x)| = \infty.$$

Fix $C > 0$ and define U as the set of all controls $u \in U_C^1$ whose corresponding trajectory of (1.1) satisfies $|x(u)(t)| \leq R$ for all t .

Then, there exists a constant K such that

$$\|\dot{u}\|_2 \leq \sqrt{K} \quad \forall u \in U.$$

COROLLARY 4.1. *Let X be the set of solutions of (1.1) corresponding to the controls belonging to U , defined as in Theorem 4.1, and let \bar{X} denote its closure in $L^1([0, T], R^n)$. If h satisfies (4.4), then*

$$\bar{X} \subset \tilde{X}_K,$$

for a suitable constant K . In particular, X is precompact in $L^1([0, T], R^n)$.

Proof of Theorem 4.1. By (4.4) and by the compactness of $B[0; R]$, there exists a $\tau > 0$ and an $\alpha \geq 1$ such that

$$(4.5) \quad R + 1 \leq |(\exp \tau h)(x)| \leq R + \alpha,$$

for every $x \in B[0; R]$.

Set

$$M_f = \max_{|x| \leq R + \alpha} |f(x)| \quad M_g = \max_{|x| \leq R + \alpha} |g(x)|$$

and let λ be the Lipschitz constant of h on $B[0; R + \alpha]$. Moreover, call N the smallest integer such that

$$(4.6) \quad e^{\lambda\tau}(TM_f + CM_g) < N.$$

We claim that

$$(4.7) \quad \|\dot{u}\|_{L^2}^2 < N\tau,$$

for every $u \in U$. Assume, on the contrary, that (4.7) does not hold for some control $u^* \in U$. Then, there exist $N + 1$ instants $0 = t_0 < t_1 < t_N \leq T$ such that

$$(4.8) \quad \int_{t_{i-1}}^{t_i} \dot{u}^{*2}(s) \, ds = \tau,$$

for every $i = 1, \dots, N$.

Denoting by $y_i : [t_{i-1}, t_i] \mapsto R^n$ the solution of the Cauchy problem

$$\begin{cases} \dot{y} = h(y)(\dot{u}^*)^2 \\ y_i(t_{i-1}) = x(u^*, t_{i-1}), \end{cases}$$

Gronwall's lemma yields

$$(4.9) \quad |x(u^*, t_i) - y_i(t_i)| \leq e^{\lambda\tau} \int_{t_{i-1}}^{t_i} |p(t)| dt,$$

for every $i = 1, \dots, N$, where

$$p(t) = f(x(u^*, t)) + g(x(u^*, t))\dot{u}^*(t).$$

Since $x(u^*, t) < R$ for every t , from (4.5), (4.7), and (4.9) we obtain

$$\int_{t_{i-1}}^{t_i} |p(t)| dt \geq e^{-\lambda\tau} |x(u^*, t_i) - y_i(t_i)| \geq e^{-\lambda\tau},$$

for every $i = 1, \dots, N$. Hence, we have

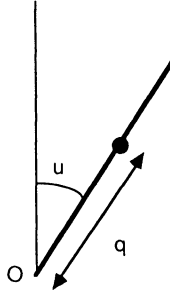
$$TM_f + CM_g \geq \int_0^T |p(t)| dt = \sum_{i=1}^N \int_{t_{i-1}}^{t_i} |p(t)| dt \geq Ne^{-\lambda\tau}.$$

This contradicts the inequality (4.6), proving the theorem with $K = N\tau$. \square

5. An application to Lagrangian mechanics. The study of equations of the form (1.1) is primarily motivated by applications (see [6]–[10], [14], [15]) to Lagrangian systems in classical mechanics. In fact, given a Lagrangian system Σ , locally parametrized by coordinates (q_i, u_α) , one can superimpose new time-dependent (holonomic) constraints by assigning the motion $u_\alpha(\cdot)$ of the last M coordinates u_α —the latter to be thought as controls. The right-hand sides of the resulting equations for the coordinates q_i and their conjugate momenta p_i contain the derivative \dot{u}_α of the controls. Typically, this functional dependence on the \dot{u}_α is quadratic, because of the Riemannian structure yielded by the kinetic energy on the state-space. In certain cases, however, it is possible to choose coordinates (q_i, u_α) (M -fit coordinates), which guarantee that the u_α appear only linearly in the equations for the q_i and the p_i (see, e.g., [8], [14], [15]). Because of the lack of an analytical theory for the quadratic case, all previous applications were confined to systems for which this special choice of the coordinates is possible ([6]–[10]).

In this section, we present an example concerning a simple Lagrangian system Σ , where the right-hand side of the dynamical equations depends on the square of the derivative of the control, regardless of the choice of the local coordinates (q_i, u_α) . In particular, we shall use some results of the previous sections to study an optimal control problem for the system Σ .

Let us consider a mechanical system Σ formed by a mass concentrated at a point P , sliding without friction along a rectilinear guide, which rotates on a vertical plane

FIG. 1. *Pendulum with variable length.*

around a fixed point O . Let q denote the distance $|P-O|$ and let u measure (clockwise) the angle formed by the ascending vertical and the segment $P-O$ (see Fig. 1).

If the active forces acting on the particle P reduce to the gravitational force, then, up to a rescaling of the physical quantities, the motion of P is governed by the following differential equations:

$$(5.1) \quad \begin{cases} \dot{q} = p \\ \dot{p} = -\cos u + q\dot{u}^2. \end{cases}$$

Let the initial conditions for q, p and u be given by

$$(5.2) \quad \begin{cases} q(0) = q_0 \\ p(0) = p_0 \\ u(0) = 0. \end{cases}$$

In connection with the control system (5.1)–(5.2), consider the optimization problem

$$(5.3) \quad \inf \left\{ \int_0^T [p^2(t) + (q(t) - \bar{q})^2] ds, \quad u \in \mathcal{U}_K \right\},$$

where

$$\mathcal{U}_K = \{u | u \in W^{1,2}([0, T], R), \quad \|\dot{u}\|_{L^2}^2 \leq K\}.$$

In (5.2) and (5.3), we assume $q_0, \bar{q} > 0$, $p_0 \in R$. Observe that the inequality $\|\dot{u}\|_2^2 \leq K$ can be thought as an a priori bound on the time average of the *kinetic energy of the guide*.

In order to apply the results of the previous sections to our problem, let us introduce the variables x_1, x_2, x_3, x_4 by setting

$$x_1 = q, \quad x_2 = p, \quad x_3 = u, \quad x_4 = \int_0^t [(x_2(s))^2 + (x_1(s) - \bar{q})^2] ds.$$

Then the system (5.1)–(5.2) is replaced by

$$(5.4) \quad \begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\cos x_3 + x_1 \dot{u}^2 \\ \dot{x}_3 = \dot{u} \\ \dot{x}_4 = (x_2)^2 + (x_1 - \bar{q})^2 \\ (x_1, x_2, x_3, x_4)(0) = (q_0, p_0, 0, 0), \end{cases}$$

and our variational problem can be written as

$$(\mathcal{P}) \quad \min\{x_4(T), \quad u \in \mathcal{U}_K\},$$

with $U_K \subset W^{1,2}$ as in §3. On the basis of the results of the previous sections we can associate to (5.4) the impulsive system

$$(5.5) \quad \begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\cos x_3 + x_1 \dot{w} \\ \dot{x}_3 = v \\ \dot{x}_4 = (x_2)^2 + (x_1 - \bar{q})^2 \\ (x_1, x_2, x_3, x_4)(0) = (\bar{q}, \bar{p}, 0, 0), \end{cases}$$

where $(v, w) \in \tilde{U}_K$, also defined in §3.

By the compactness of the set reachable by solutions of (5.5) (see Cor. 3.1), for any $K > 0$ we obtain the existence of an optimal control pair for the extended problem

$$(\mathcal{P}') \quad \min \{x_4(T), \quad (v, w) \in \tilde{U}_K\},$$

where now $x_4(\cdot)$ denotes the fourth component of the solution of (5.5). In order to explicitly compute the optimal solution, when K is sufficiently large, we first write the Euler necessary conditions for the variational problem

$$(5.6) \quad \min \int_0^T [(q(t) - \bar{q})^2 + \dot{q}^2(t)] dt, \quad q(0) = q_0.$$

Solving the two-point boundary value problem

$$\ddot{q}(t) = q(t) - \bar{q}, \quad q(0) = q_0, \quad \dot{q}(T) = 0,$$

one finds the unique optimal solution of (5.6)

$$(5.7) \quad q(t) = \bar{q} + \left(\frac{q_0 - \bar{q}}{\cosh(T)} \right) \cosh(t - T).$$

Therefore, if K is sufficiently large and

$$(5.8) \quad \dot{q}(0+) = \left(\frac{q_0 - \bar{q}}{\cosh(T)} \right) \sinh(-T) \geq p_0,$$

then the trajectory

$$(5.9) \quad (x_1, x_2, x_3, x_4)(t) = \left(q(t), \dot{q}(t), 0, \int_0^t [(q(s) - \bar{q})^2 + \dot{q}^2(s)] ds \right)$$

is an optimal one for the generalized system (5.5).

This trajectory corresponds to the control pair (v, w) , with $v(t) \equiv 0$, $w(0) = 0$ and

$$(5.10) \quad w(t) = \frac{\dot{q}(0+) - p_0}{q_0} + \int_0^t \frac{1 + \ddot{q}(s)}{q(s)} ds \quad \forall t > 0.$$

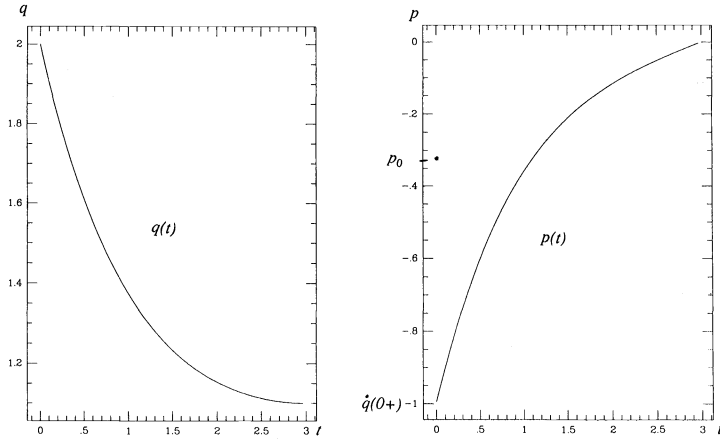


FIG. 2. The optimal trajectories $q(t)$ and $p(t) [= \dot{q}(t)]$.

More precisely, (5.9) yields an optimal trajectory in the class \tilde{X}_K provided that (5.8) holds together with the bounds

$$(5.11) \quad \dot{\omega}(t) \geq 0 \quad \forall t > 0$$

and

$$w(T) = \frac{\dot{q}(0+) - p_0}{p_0} + \int_0^T \frac{1 + \ddot{q}(s)}{q(s)} ds \leq K.$$

For instance, (5.11) is automatically satisfied when $q_0 > \bar{q}$ (see Fig. 2).

We remark that, if (5.8) fails, then (5.9) is no longer an admissible trajectory of (5.5). This happens because the centrifugal force $x_1 \dot{w} = x_1 \dot{u}^2$ can only assume nonnegative values. Hence, a jump in w can take \dot{q} from p_0 to the optimal initial value $\dot{q}(0+)$ only when (5.8) holds.

Observe that, whenever the inequality in (5.8) is strict, no optimal control can exist for the original problem (\mathcal{P}) . Indeed, for every control $u \in W^{1,2}$, the corresponding solution $q = q(u)$ of (5.1) has an absolutely continuous derivative, satisfying

$$\lim_{t \rightarrow 0+} \dot{q}(t) = p_0,$$

hence, it cannot coincide with the unique optimal solution of (5.6).

On the other hand, if $\dot{q}(0+) = p_0$, then several optimal controls exist for the original problem (\mathcal{P}) . To construct such controls, define the function $\dot{q}(\cdot)$ as in (5.7) and let u be a solution to the multivalued differential equation

$$\dot{u}(t) = \pm \sqrt{\frac{\dot{q}(t) + \cos u}{q(t)}}, \quad u(0) = 0,$$

defined on $[0, T]$. The corresponding trajectory of (5.4) then satisfies

$$(x_1, x_2, x_3, x_4)(t) = \left(q(t), \dot{q}(t), u(t), \int_0^t [(q(s) - \bar{q})^2 + \dot{q}^2(s)] ds \right),$$

and therefore, is optimal.

REFERENCES

- [1] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer, Berlin, 1984.
- [2] A. BRESSAN, *On differential systems with impulsive controls*, Rend. Sem. Mat. Univ. Padova, 78 (1987), pp. 227–236.
- [3] A. BRESSAN AND F. RAMPAZZO, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.
- [4] ———, *Impulsive control systems without commutative assumptions*, preprint 147M SISSA, Trieste, Italy, 1990.
- [5] ———, *On differential systems with vector-valued impulsive controls*. Bull. Univ. Mat. Ital. B(7), 3 (1988), pp. 641–656.
- [6] A. BRESSAN, *On control theory and its applications to certain problems for Lagrangian systems. On hyperimpulsive motions for these (I, II)*. Atti Accad. Naz. Lincei Rend. Cl. Sc. Fis. Mat. Natur. (8), 82 (1988), pp. 91–118.
- [7] ———, *On control theory and its applications to certain problems for Lagrangian systems. On hyperimpulsive motions for these (III)*. Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 82 (1988), pp. 461–471.
- [8] ———, *Hyperimpulsive motions and controllizable coordinates for Lagrangean systems*. Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. (8), 19 (1989).
- [9] ———, *On some control problem concerning the ski and the swing*, Atti Accad. Naz. Lincei, Mem. Cl. Sci. Fis. Mat. Natur. (9), 1 (1991), pp. 149–196.
- [10] ———, *On some recent results in control theory, for their applications to Lagrangean systems*, Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. (8), 19 (1989).
- [11] H. BREZIS, *Analyse fonctionnelle*, Masson, Paris, 1987.
- [12] G. DAL MASO AND F. RAMPAZZO, *On systems of ordinary differential equations with measures as controls*, Differential and Integral Equations, 4 (1991), pp. 739–765.
- [13] M. A. KRASNOSEL'SKII AND A. V. POKROVSKII, *Vibrostable differential equations with a continuous right-hand side*. Proc. Moscow Math. Soc., 27 (1972), pp. 93–113.
- [14] F. RAMPAZZO, *On Lagrangean systems with some coordinates as controls*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 82 (1988), pp. 685–695.
- [15] ———, *On the Riemannian structure of a Lagrangian system and the problem of adding time-dependent constraints as controls*, European J. Mech. A Solids, 10 (1991), pp. 405–431.
- [16] ———, *Optimal impulsive controls with a constraint on the total variation*, in New Trends in Systems Theory, G. Conte, A. M. Perdon, B. F. Wyman, eds., Series Progress in Systems and Control Theory, Birkhauser, Boston, 7 (1991), pp. 606–613.
- [17] A. V. SARYCHEV, *Nonlinear systems with impulsive and generalized function controls*, Proc. Conf. on Nonlinear Synthesis, Sopron, Hungary, 1989.
- [18] H. J. SUSSMANN, *On the gap between deterministic and stochastic ordinary differential equations*, Ann. Probab., 6 (1978), pp. 17–41.

IDENTIFICATION OF THE COEFFICIENT IN ELLIPTIC EQUATIONS*

ROBERT ACAR[†]

Abstract. This paper seeks to identify the space-distributed diffusion coefficient a in the steady-state diffusion equation $-\operatorname{div}(a \operatorname{grad} u) = f$ in a bounded region of the real plane; u , f , and the flux at the boundary are observed. This problem is solved by a least square method, choosing for the cost function the “equation error” defined in a natural way by the Riesz representation theorem. The method is stable if a smoothing term is added to the cost function (Tikhonov regularisation); this paper shows that selection strategies of the smoothing coefficient exist that insure convergence under the only condition that a be sufficiently smooth. The case is also considered where a is piecewise smooth and it is shown how to extend the previous formulation. In both cases computational results are presented.

Key words. parameter estimation, inverse problems

AMS subject classifications. 35, 49

1. Introduction. The steady-state diffusion equation in a transmissive two-dimensional medium, using a familiar notation, is

$$(1) \quad -\nabla(a(x) \nabla u(x)) = f(x).$$

This holds in a domain Ω of the plane; $a(x)$ is the permeability coefficient, $u(x)$ the fluid potential, and f the sink distribution. We seek to solve the problem: *Given observations of f and ∇u , and some additional information involving a , determine the permeability $a(x)$.* This problem, as well as the one arising from the time-dependent diffusion equation, has been studied extensively:

$$(2) \quad c(x) \frac{\partial u}{\partial t} - \nabla(a \nabla u) = f(t, x).$$

Indeed, (2) models groundwater flow, with a being the space-distributed permeability and u the piezometric head. The domain is then three-dimensional, but in most situations the permeability varies very slowly, or not at all, with depth, so that (assuming that sinks are placed vertically) the flow may be viewed as two dimensional.

Since the medium is transmissive everywhere, there holds $a(x) \geq c_0 > 0$ (even though we may not know c_0 a priori). As an equation in u , (1) is then seen to be elliptic. The *forward problem* is regular in the sense that small variations of a in L^∞ induce small variations of u in H^1 . However, since we are interested in the *inverse problem* (that of recovering a from ∇u), we need to know that large variations in a induce large variations in ∇u (so that small variations in ∇u could only be caused by small variations in a , boding well for the sensitivity of the measurements ∇u). Unfortunately, this is not the case; we can find examples of sequences (a_n) that remain

*Received by the editors January 16, 1989; accepted for publication (in revised form) April 3, 1992.

[†]Eastern Montana College, Box 542, Billings, Montana 59101. The work of this author was supported in part by the Air Force Office of Scientific Research grant 85-0263 and the Graduate Resources Committee and the Mathematics Research Center at the University of Wisconsin–Madison.

uniformly distant from a certain a_0 in the L^∞ norm, and yet such that the gradient ∇u_n of the corresponding solution tends to ∇u_0 in L^2 . This says that the inverse problem is ill posed with respect to the above topologies. Of importance in such examples is the oscillatory behaviour of a_n in one of the variables, similar to the example in one dimension given by Murat (see [L]). These examples show that it is hopeless to try solving (1) unless we assume that the oscillations in a remain bounded; in other words, a successful numerical scheme producing iterates that have a limit in L^∞ must enforce some “compactification” which, in effect, penalises any oscillatory behaviour.

There have been two main thrusts in attacking the inverse problem. The first thrust reformulates the problem as an optimisation problem and relies on some iterative procedure to solve it. The second thrust uses the observation that (1) or (2) is a hyperbolic first-order partial differential equation (PDE) in a and uses direct methods to solve it. Although the first methods may have some smoothing embedded in them, direct methods are more vulnerable to the illposedness and must include some auxiliary smoothing device.

In [R1], Richter uses a direct method to solve the steady-state identification problem (using data on a on the inflow boundary). Making some assumptions on u (such as $\inf_\Omega \max(|\nabla u|, \Delta u) > 0$), he derives a continuous dependence result on the hyperbolic problem, and formulates favorable “test conditions” of u (i.e., for which no Cauchy data is needed on a). In [R2], Richter applies the previous ideas to the discretised case, using a finite-difference scheme explicit upwind along the characteristics (that happen to be the curves of steepest ascent of u); the previous estimates on $|a|_\infty$ are carried over to the discrete setup. Falk ([F]) minimises an observation error criterion $J(u, a)$ given an observation z on u ; the admissible u ’s obey a Neumann condition, and are normalised so that $\int_\Omega (u - z) = 0$. u is computed from a by writing (1) in weak form (see §2). Then under assumptions of smoothness of u , and of the boundary piece where $\partial u / \partial n > 0$, and an assumption on the field ∇u (its direction is contained in a proper cone), Falk is able to estimate the L^2 -norm of the error in a in terms of the L^2 -norm of the error in u (and of the size of the grid). Kravaris and Seinfeld ([KS]) seek to minimise a similar cost function, and they add a smoothing term that includes the H^2 -norm of a (Tikhonov regularisation) to insure convergence and stability; they show how this may be applied to coefficient identification in a class of second-order linear parabolic systems. Hoffman and Sprekels ([HS]) solve the problem by embedding it in a family of time-dependent parabolic equations and regarding it as the steady-state limit. In [A], Alessandrini examines the identification problem with boundary data on a , and establishes convergence of an elliptic regularisation (singular perturbation) algorithm under regularity hypotheses in a with a stability result with respect to the u data. Chicone and Gerlach ([CG]) define pointwise uniqueness of $a(x)$, and characterise the region of uniqueness in terms of “nonwandering” sets of the dynamical system associated with the gradient field of u . Closest to the spirit of the present work, Kohn and Lowe ([KL]) investigate methods that handle different aspects of the equation error, some including a regularising term. By choosing weights related to the mesh size, they obtain convergence rates under the assumption that $\|u^m - u_*\|$ is $O(h^\alpha)$ where $\alpha > 1$, u_* is the true pressure u^m is the measured pressure, and h the mesh size.

Without making any claim to completeness, let us also mention [Kr], [JJ], [J], [EdM], [FP], [CDL], [Kl], [LY], [P], [CS], [BK2], [S], [Ne], [SYD], [Nu], [CY] and [YT]. For other references as well as a comprehensive discussion of the background of this problem, see [BK1].

An inherent difficulty with direct methods is the choice of the piece of the bound-

ary where we specify a . Since information is carried along the streamlines, we must specify a on a piece of the boundary where each streamline is intersected exactly once. This choice depends, in turn, on the geometry of the streamlines, which is sensitive to reading errors in ∇u . The advantage of methods which rely on error minimisation is that they easily lend themselves to overdeterminacy of the data, such as observing a all around the boundary of the domain.

We describe how to carry out the identification of a in (1) by a least square method. We cope with the illposedness of a by adding a smoothing term and assuming that a is sufficiently smooth. In §2, we derive the convergence and stability results for the equation error method. In §3, we discuss the numerical solution and present results. In §4, we show how the method can be extended to the case where a is discontinuous along fracture lines in the domain, and give computational results as well.

2. The equation error method.

2.1. Existence. If we are to solve the inverse problem (viewed as a first-order PDE in a) exactly, assuming that we have no data on a in the interior of Ω , we need to know a on the boundary. We assume in this section that a is continuous.

∇u determines the geometry of characteristics in Ω , which in turn determines whether a portion of the boundary is suitable for specifying a . Since we do not quite know which portion of Γ is a good portion to chose (not exactly knowing ∇u), we assume that we observe the flux $a(\partial u)/(\partial n)$ throughout the boundary Γ , say, $a(\partial u)/(\partial n) = g$.

Denoting $H^1(\Omega)$ by V , we may then rewrite (1) in variational form

$$(3) \quad \forall \phi \in V, \quad \int_{\Omega} a \nabla u \nabla \phi = \langle f, \phi \rangle_{V', V} + \int_{\Gamma} g \phi.$$

We see that if we take this as the effective interpretation of (1), we only need a to be in L^∞ ; so the assumption that a is continuous is sufficient.

This problem has no solution unless g is “consistent” (i.e., the whole of it consistent with any part construed as Cauchy data). This may again be asking for too much, since there are errors in observing g . We avoid this by trying to make the “equation error”

$$- \int_{\Omega} a \nabla u \nabla \phi + \langle f, \phi \rangle_{V', V} + \int_{\Gamma} g \phi$$

small, in the following sense: Note that

$$\phi \mapsto - \int_{\Omega} a \nabla u \nabla \phi + \langle f, \phi \rangle_{V', V} + \int_{\Gamma} g \phi$$

is a linear functional, continuous over V (using Cauchy–Schwarz for the first term, and continuity of the 0-trace from $H^1(\Omega)$ into $L^2(\Gamma)$ for the third term), so by the Riesz representation theorem, then

$$(4) \quad \exists v \in V: \quad \langle v, \phi \rangle_V \equiv - \int_{\Omega} a \nabla u \nabla \phi + \langle f, \phi \rangle_{V', V} + \int_{\Gamma} g \phi.$$

Then putting $J_e(a) = |v(\nabla u, a)|_V^2$, the problem at hand is to minimise $J_e(a)$.

Noting that $v(\nabla u, a)$ is an affine function of a , we see that $J_e(a)$ is convex. Furthermore, if $v(a)$ is injective, then $J_e(a)$ is strictly convex, and a minimising it

would be unique, as long as the domain of a is convex. When a is continuous, a sufficient condition that v be injective is that ∇u be zero at most on a set of measure zero: indeed, if v_1 and v_2 correspond to a_1 and a_2 , then $\langle v_2 - v_1, \phi \rangle = -\int (a_2 - a_1) \nabla u \nabla \phi$ for all ϕ in V , and if $a_2 - a_1 \neq 0$ in C^0 , then $a_2 - a_1 \neq 0$ on an open set, and the same holds for $(a_2 - a_1) \nabla u$, so choosing $\phi = u$ shows that $v_2 - v_1 \neq 0$.

But the problem of minimising J_e for a , say, in L^2 or C^0 , is not stable: *L^2 -small perturbations in ∇u may produce large perturbations in a .*

2.2. Stability. To make the minimisation problem stable, we need to introduce some compactification in a . One way to do this is to impose a priori bounds on a , as many have done in the case of the observation error method. Another useful device, which is the one we use, is *Tikhonov regularisation* (introduced by Tikhonov in conjunction with the solution of Fredholm integral equations of the first kind; see [Ti1] and [Ti2]). It consists of adding a smoothing term (the norm of a in a sufficiently high Sobolev space) to the cost function.

Hence we set to minimise the smoothed cost function

$$\begin{aligned} J_\lambda(\nabla u, a) &= |v(\nabla u, a)|_V^2 + \lambda |a|_{\mathcal{R}}^2, \\ &=: J_e(\nabla u, a) + \lambda J_s(a). \end{aligned}$$

Among the Sobolev spaces $W^{n,2}$ (i.e., with derivatives up to order n in L^2) embedded in $C^0(\Omega)$, $W^{2,2}$ (also denoted H^2) is the one of lowest order. Therefore, take the set of admissible a 's to be \mathcal{R} , where

$$\mathcal{R} = \{ a \in H^2(\Omega) : a \frac{\partial u}{\partial n} = g \text{ on } \Gamma \}.$$

This choice insures the existence of a global minimum of $J_\lambda(\nabla u, \cdot)$. Indeed, if $m = \inf \{ J_\lambda(\nabla u, a) : a \in \mathcal{R} \}$, there exists (a_n) in $\mathcal{R} : \lim \downarrow J_\lambda(\nabla u, a_n) = m$. Since $J_\lambda(\nabla u, a_n)$ decreases, the sequence (a_n) remains bounded in \mathcal{R}

$$\begin{aligned} \forall n, \quad |a_n|_{\mathcal{R}}^2 &\leq \frac{1}{\lambda} J_\lambda(\nabla u, a_n), \\ &\leq \frac{1}{\lambda} J_\lambda(\nabla u, a_1). \end{aligned}$$

Then there is a subsequence $(a_{n,k})$ weakly converging to some $\bar{a} \in \mathcal{R}$, and convergence also holds in the C^0 norm. It follows then that \bar{a} yields the global minimum of $J_\lambda(\nabla u, \cdot)$ from the following two facts: (i) $a \mapsto |a|_{\mathcal{R}}^2$ is weakly lower semicontinuous, and (ii) $a \mapsto |v(\nabla u, a)|_V^2$ is continuous. Indeed, a perturbation δa of a corresponds to the perturbation δv solving

$$\forall \phi \in V, \quad \langle \delta v, \phi \rangle_V = - \int \delta a \nabla u \nabla \phi,$$

so that

$$\begin{aligned} |\langle \delta v, \phi \rangle_V| &\leq |\delta a|_\infty |\nabla u|_{L^2} |\nabla \phi|_{L^2}, \\ &\leq |\delta a|_\infty |\nabla u|_{L^2} |\phi|_V \end{aligned}$$

implying that

$$|\delta v|_V \leq |\nabla u|_{L^2} |\delta a|_\infty,$$

and then,

$$\delta J_e = 2 \langle v, \delta v \rangle_V + |\delta v|_V^2,$$

implying that

$$|\delta J_e| \leq 2|v(\nabla u, a)|_V \cdot |\delta v|_V + |\delta v|_V^2,$$

so $|\delta J_e|$ is, in turn, bounded in terms of $|\delta a|_\infty$.

Also, \bar{a} is unique by strict convexity of $J_\lambda(\nabla u, \cdot)$ (J_e is convex, and the \mathcal{R} -norm is strictly convex). We denote the global minimum of $J_\lambda(\nabla u, \cdot)$ by a_λ .

Note that we did not require a in \mathcal{R} to be bounded away from zero; this is justified by the next theorem, where, for ε small enough, a_λ will be bounded away from zero if a_* is also bounded away.

THEOREM 1 (Stability). *Assume that (u_*, a_*) is a solution pair of (3) (so that $v(\nabla u_*, a_*) = 0$), and that $v(\nabla u_*, \cdot)$ is injective (say, $\nabla u_* = 0$ at most on a set of measure 0). Then*

$$\begin{aligned} \forall \varepsilon > 0, \quad \exists \delta_0 > 0 : \quad \forall \delta \leq \delta_0, \quad \forall \lambda \in [l(\delta), r(\delta)], \\ |\nabla u - \nabla u_*|_{L^2} \leq \delta \quad \Rightarrow \quad |a_\lambda - a_*|_\infty \leq \varepsilon, \end{aligned}$$

where l, r are functions satisfying

- (i) $\exists C > 0 : \quad \delta^2 \leq Cl(\delta),$
- (ii) $r(\delta) > l(\delta),$ and $\lim r(\delta) = 0$ as $\delta \rightarrow 0$.

(Recall from a previous observation that a_* is then unique.) The theorem states that a_λ converges to a_* as λ and the error on ∇u_* go to zero, but they cannot be arbitrarily made small; (i) states that λ should not converge to zero faster than the square of δ , otherwise we lose our grip on a_λ . The proof uses the following observations.

Remark 1. For all admissible a , $|v(\nabla u, a) - v(\nabla w, a)|_V \leq |a|_\infty |\nabla u - \nabla w|_{L^2}$. Indeed, for all $\phi \in V$,

$$\langle v, \phi \rangle_V = - \int a (\nabla u - \nabla w) \nabla \phi.$$

Remark 2. If $|\nabla u - \nabla u_*|_{L^2} \leq \delta \leq \delta_0$ and $\lambda \in [l(\delta), r(\delta)]$ (where $l(\delta), r(\delta)$ are as above), then a_λ stays uniformly bounded in \mathcal{R} :

$$\begin{aligned} \lambda |a_\lambda|_{\mathcal{R}}^2 &\leq J_\lambda(\nabla u, a_\lambda) \\ &\leq J_\lambda(\nabla u, a_*) \\ &= |v(\nabla u, a_*)|_V^2 + \lambda |a_*|_{\mathcal{R}}^2 \\ &\leq |a_*|_\infty^2 \delta^2 + \lambda |a_*|_{\mathcal{R}}^2 \\ &\leq |a_*|_\infty^2 C \lambda + \lambda |a_*|_{\mathcal{R}}^2; \end{aligned}$$

(this is where we needed (i))

$$\Rightarrow |a_\lambda|_{\mathcal{R}}^2 \leq C |a_*|_\infty^2 + |a_*|_{\mathcal{R}}^2.$$

Proof of Theorem 1. Let us assume that the claim is not true. Then

$$\begin{aligned} \exists \varepsilon > 0 : \quad \forall n, \quad \exists \delta_n \leq 1/n, \quad \exists \lambda_n \in [l(\delta_n), r(\delta_n)], \quad \exists u_n : \\ |\nabla u_n - \nabla u_*| \leq \delta_n \quad \&\quad |a_{\lambda_n} - a_*|_\infty > \varepsilon. \end{aligned}$$

By Remark 2, $a_{\lambda_n} \rightharpoonup \tilde{a}$ (weakly in \mathcal{R}) along a subset \mathbb{M} of \mathbb{N} , implying that $a_{\lambda_n} \rightarrow \tilde{a}$ in C^0 . For any λ_n , denoting a_{λ_n} by a_n for short,

$$|v(\nabla u_*, a_n)|_V \leq |v(\nabla u_n, a_n)|_V + |v(\nabla u_*, a_n) - v(\nabla u_n, a_n)|_V.$$

By Remark 1, the second term is smaller than $|a_n|_\infty \delta_n$, and $|a_n|_\infty$ remains bounded by Remark 2. For the first term,

$$\begin{aligned} |v(\nabla u_n, a_n)|_V^2 &\leq J_{\lambda_n}(\nabla u_n, a_n) \\ &\leq J_{\lambda_n}(\nabla u_n, a_*) \\ &\leq |a_*|_\infty^2 \delta_n^2 + \lambda_n |a_*|_{\mathcal{R}}^2, \end{aligned}$$

so that $\lim |v(\nabla u_*, a_n)|_V = 0$ as $n \rightarrow \infty$, $n \in \mathbb{M}$, contradicting $\lim |v(\nabla u_*, a_n)|_V = |v(\nabla u_*, \tilde{a})|_V \neq 0$, since $\tilde{a} \neq a_*$.

2.3. Parameter selection strategies. For Theorem 1 to be fully relevant, it should allow us to prescribe rules for choosing the smoothing coefficient λ , given a bound on the error $|\nabla u - \nabla u_*|_{L^2}$ and possibly information on a_* . Two such rules are discussed in [KS] in the case of the observation-error method. We briefly indicate how these rules also be applied to the equation-error method.

One way (see, also, [Mi]) is as follows: If K is a known upper bound on $|a_*|_\infty$, E on $|a_*|_{\mathcal{R}}$, and δ on $|\nabla u - \nabla u_*|_{L^2}$, then choose $l(\delta)$ and $r(\delta)$ proportional to $K^2\delta^2/E^2$.

Indeed, if, say,

$$l(\delta) = \gamma K^2 \delta^2 / E^2, \quad r(\delta) = \Gamma K^2 \delta^2 / E^2,$$

then (i) and (ii) of Theorem 1 hold, and

$$\begin{aligned} J_\lambda(\nabla u, a_\lambda) &\leq J_\lambda(\nabla u, a_*) \\ &\leq |a_*|_\infty^2 \delta^2 + \lambda |a_*|_{\mathcal{R}}^2 \\ &\leq K^2 \delta^2 + \Gamma K^2 \delta^2, \end{aligned}$$

so that

$$|v(\nabla u, a_\lambda)|_V \leq \sqrt{1 + \Gamma} K \delta$$

and

$$|a_\lambda|_{\mathcal{R}} \leq \sqrt{\frac{1 + \Gamma}{\gamma}} E.$$

To choose γ and Γ , keeping in mind that the a priori bounds on $|v(\nabla u, a_*)|_V$ and $|a_\lambda|_{\mathcal{R}}$ are $K\delta$ and E , we may ask that the larger of the factors $\sqrt{1 + \Gamma}$ and $\sqrt{\frac{1 + \Gamma}{\gamma}}$ be as small as possible; i.e., solve

$$\text{minimise } \left\{ \max \left(\sqrt{1 + \Gamma}, \sqrt{\frac{1 + \Gamma}{\gamma}} \right) : 0 < \gamma < \Gamma \right\}.$$

The solution is easily seen to be $\gamma = \Gamma = 1$, which is Miller's choice [Mi].

Another method of choosing λ , suggested by Tikhonov and Arsenin [TA], is this: if $\sqrt{\delta}$ is a bound on $|\nabla u - \nabla u_*|_{L^2}$, choose λ such that the corresponding a_λ satisfies

$$|v(\nabla u, a_\lambda)|_V^2 = \delta.$$

It will follow that $\lim |a_\lambda - a_*|_\infty = 0$ as $\delta \rightarrow 0$; first, we must show the existence of such a λ .

Some notation is given as follows:

$$\begin{aligned} J_\lambda(\nabla u, a) &:= |v(\nabla u, a)|_V^2 + \lambda |a|_{\mathcal{R}}^2 = J_e(\nabla u, a) + \lambda J_s(a); \\ a_{\min} &:= \text{the min-norm element of } \mathcal{R}; \\ \delta_{\max} &:= J_e(\nabla u, a_{\min}); \\ \delta_{\min} &:= \inf \{ J_e(a) : a \in \mathcal{R} \}. \end{aligned}$$

(a) Given ∇u and the corresponding δ_{\min} and δ_{\max} , then if $\delta_{\min} < \delta < \delta_{\max}$, there exists λ such that the minimiser a_λ of $J_\lambda(\nabla u, a)$ over \mathcal{R} satisfies $J_e(a_\lambda) = \delta$. We sketch the following proof.

(i) The map $\lambda \mapsto a_\lambda$ is continuous for $\lambda > 0$: if $\lambda \leq \mu$, it is easily seen that $J_S(a_\lambda) \geq J(a_\mu)$ and that $J_e(a_\lambda) \leq J_e(a_\mu)$. We may then write

$$\begin{aligned} J_e(a_\lambda) + \lambda J_S(a_\lambda) &\leq J_e(a_\mu) + \lambda J_S(a_\mu) \\ &\leq J_e(a_\mu) + \mu J_S(a_\mu) \\ &\leq J_e(a_\lambda) + \mu J_S(a_\lambda). \end{aligned}$$

Assume for now that λ is fixed and μ approaches λ from above; then $J_\lambda(a_\mu) - J_\lambda(a_\lambda) \leq J_\mu(a_\lambda) - J_\lambda(a_\lambda)$. Since J_λ has bounded lower level sets, we see that a_μ is bounded in \mathcal{R} , and weakly converges to a_λ . The case where μ is fixed and λ approaches μ from below is similar.

(ii) For some value of λ , $J_\lambda(a_\lambda) < \delta$: choose indeed λ such that $J_\lambda(a_0) < \delta$, where $J_e(a_0) = \delta_{\min}$.

(iii) $\lim_{\lambda \rightarrow \infty} J_\lambda(a_\lambda) = \delta_{\max}$: since $|a_\lambda|_{\mathcal{R}}^2 \leq (J_S + \frac{1}{\lambda} J_e)(a_\lambda) \leq (J_S + \frac{1}{\lambda} J_e)(a_{\min})$, $|a_\lambda|_{\mathcal{R}}^2$ will be bounded for $\lambda \geq 1$. So $J_e(a_\lambda)$ stays bounded since J_e is weakly continuous, and any weakly convergent subsequence (as $\lambda \rightarrow \infty$) must have a_{\min} as limit: if $a_\lambda \rightharpoonup \bar{a}$, then $J_S(a_{\min}) \geq \liminf J_S(a_\lambda) \geq J_S(\bar{a})$ since $(J_S + \frac{1}{\lambda} J_e)(a_{\min}) \geq (J_S + \frac{1}{\lambda} J_e)(a_\lambda)$.

We conclude by the intermediate value theorem.

(b) For λ such as in (a), a_λ yields the minimum of J_S over $\{J_e \leq \delta\}$. Indeed, if $J_e(a) \leq \delta$ and $J_S(a) \leq J_S(a_\lambda)$, then $(J_e + \lambda J_S)(a) \leq \delta + \lambda J_S(a_\lambda) = (J_e + \lambda J_S)(a_\lambda)$, implying that $a = a_\lambda$.

The following convergence theorem then justifies using the Tikhonov parameter selection rule.

THEOREM 2. Assume that $\exists! a_* \in \mathcal{R} : v(\nabla u_*, a_*) = 0$. Then

$$\begin{aligned} \forall \varepsilon > 0, \quad \exists \delta_0 > 0 : \forall \delta \leq \delta_0, \\ |\nabla u - \nabla u_*|_{L^2} < \alpha \sqrt{\delta} \implies |a_{\lambda(\delta)} - a_*|_\infty \leq \varepsilon, \end{aligned}$$

where $\alpha = 1/|a_*|_\infty$, and $\lambda(\delta)$ is a smoothing parameter given by (a) above; i.e., such that $J_e(a_{\lambda(\delta)}) = \delta$.

By the discussion above, $\lambda(\delta)$ exists, for if $|\nabla u - \nabla u_*|_{L^2} < \alpha \sqrt{\delta}$, then $\min \{J_e(\nabla u, a) : a \in \mathcal{R}\} < \delta$, because

$$\begin{aligned} J_e(\nabla u, a_*) &= |v(\nabla u, a_*)|_V^2 \\ &\leq |a_*|_\infty^2 |\nabla u - \nabla u_*|_{L^2}^2 \\ &< \delta \end{aligned}$$

(i.e., in our notation, $\delta > \delta_{\min}$). Recall also that a_* , if it exists, is unique.

Proof. Let $\tilde{A} := \{a \in \mathcal{R} : |a|_{\mathcal{R}} \leq |a_*|_{\mathcal{R}}\}$. We assume that the claim is not true, then

$$\begin{aligned} \exists \varepsilon > 0 : \quad \forall n, \quad \exists \delta_n \leq 1/n, \\ \exists u_n : \quad |\nabla u - \nabla u_*|_{L^2} \leq \alpha \sqrt{\delta_n}, \\ \exists \lambda_n, \end{aligned}$$

such that

$$\begin{aligned} |v(\nabla u_n, a_{\lambda_n})|_V^2 &= \delta_n, \\ |a_{\lambda_n} - a_*|_\infty &> \varepsilon. \end{aligned}$$

a_{λ_n} minimises J_S over $\{a \in \mathcal{R} : |v(\nabla u_n, a)|_V^2 \leq \delta_n\}$. However,

$$\begin{aligned} |v(\nabla u_n, a_*)|_V^2 &\leq |a_*|_\infty^2 |\nabla u_n - \nabla u_*|_{L^2}^2 \\ &\leq \delta_n, \end{aligned}$$

so a_* belongs to the set above; hence $J_S(a_{\lambda_n}) \leq J_S(a_*)$ implying that the a_{λ_n} all stay in \tilde{A} . So along some subsequence, $a_{\lambda_n} \rightarrow \hat{a}$ in C^0 . Then

$$|v(\nabla u_*, a_{\lambda_n})|_V \leq |v(\nabla u_n, a_{\lambda_n})|_V + |v(\nabla u_*, a_{\lambda_n}) - v(\nabla u_n, a_{\lambda_n})|_V,$$

so the left-hand side tends to zero as $n \rightarrow \infty$. But it also tends to $|v(\nabla u_*, \hat{a})|_V$, which cannot be zero since $\hat{a} \neq a_*$ ($|\hat{a} - a_*|_\infty \geq \varepsilon$). Finally, if a_1 minimises J_{λ_1} and a_2 minimises J_{λ_2} , then $\lambda_1 < \lambda_2 \Rightarrow J_e(a_{\lambda_1}) \leq J_e(a_{\lambda_2})$.

Consequence. To find $\lambda(\delta)$, use an interval-halving algorithm.

Remark 3. The question of selection strategies is of obvious theoretical interest; however, in practice, the cost of computations precludes finding the smoothing parameter that exactly satisfies the criterion. For one thing, we should not expect to have a real bound, but rather a rough estimate on the error δ . Other fairly general selection strategies and their extensions are discussed in great detail by Morozov [Mo].

3. Numerical implementation.

3.1. Procedure. We describe numerical experiments with the equation-error method. We choose a , u , and f satisfying (1); assuming u , f , and the flux data on the boundary given, compute a by the equation-error method and compare it to the real a .

We use a finite-element discretisation with an uniform rectangular grid on the square $\Omega = (-1, 1) \times (-1, 1)$. We use the same grid for the finite-element approximations of v and a . They are tensor products of C^0 (respectively, C^1) univariate splines; the latter are the first and second Hermite cubics. Their tensor products assume, respectively, one (for C^0) or four (for C^1) shapes.

Call V_h and A_h the finite-dimensional discretisations of the admissible spaces for v and a . The “degrees of freedom” are: for v^h , the value of the function; for a^h , the value of the function and of its first derivatives and mixed second derivatives at the nodes.

The discrete version of (4) is

$$(5) \quad \forall \phi^h \in V_h, \quad \langle v^h, \phi^h \rangle_V = - \int a^h \nabla u \nabla \phi^h + \langle f, \phi^h \rangle_{V', V} + \int_\Gamma g \phi^h,$$

and the discrete minimisation is over the quantity $J_\lambda^h(a^h) := J_e^h(\nabla u, a^h) + \lambda |a^h|_{\mathcal{R}}^2$ where $J_e^h(\nabla u, a^h) = |v^h(\nabla u, a^h)|_{H^1}^2$. Since this is a quadratic in a^h , we use the conjugate gradient method. The initial point is taken to be a starting guess.

3.2. Convergence. To discuss convergence of the “discrete” solution (solution of the discretised problem) to that of the continuous problem, we introduce more notation: recall that $v \in V (= H^1(\Omega))$, $a \in \mathcal{R}$. By discretising, we are replacing V and \mathcal{R} by finite-dimensional subspaces V_h and \mathcal{R}_k provided with the induced norms. For a in \mathcal{R} , $v^h(\nabla u, a)$ solves

$$\langle v^h, \phi \rangle_V = - \int a \nabla u \nabla \phi + \langle f, \phi \rangle_{V', V} + \int_\Gamma g \phi, \quad \phi \in V_h,$$

and a_λ^k solves

$$\min |v^h(\nabla u, a)|_V^2 + \lambda |a|_{\mathcal{R}}^2, \quad a \in \mathcal{R}_k.$$

A consequence is that $|v^h(\nabla u, a)|_V \leq |v(\nabla u, a)|_V$.

Theorem 1 now acquires more elaborate wording.

THEOREM 1'. *Under the hypotheses of Theorem 1, and assuming that $\bigcup_h V_h$ and $\bigcup_k \mathcal{R}_k$ are dense in V and \mathcal{R} , respectively,*

$$\begin{aligned} \forall \varepsilon > 0, \quad \exists \delta_0 > 0 : \quad \forall \delta \leq \delta_0, \quad \forall \lambda \in [l(\delta), r(\delta)], \\ \exists h_0, k_0 : \quad \forall h \leq h_0, \quad \forall k \leq k_0, \\ |\nabla u - \nabla u_*|_{L^2} \leq \delta \quad \Rightarrow \quad |a_\lambda^k - a_*|_\infty \leq \varepsilon. \end{aligned}$$

The proof relies on Remark 1 following Theorem 1, and on Remark 2 where a_λ^k replaces a_λ .

Fact. If $|\nabla u - \nabla u_*|_{L^2} \leq \delta \leq \delta_0$ and $\lambda \in [l(\delta), r(\delta)]$, then a_λ^k stays bounded in \mathcal{R} as long as $k(\lambda)$ is such that $|a_*^k - a_*|_{\mathcal{R}}^2 \leq \lambda$, where a_*^k is the projection of a_* on \mathcal{R}_k . Indeed,

$$\begin{aligned} \lambda |a_\lambda^k|_{\mathcal{R}}^2 &= J_\lambda^h(\nabla u, a_\lambda^k) \\ &\leq J_\lambda^h(\nabla u, a_*^k) \\ (6) \quad &= |v^h(\nabla u, a_*^k)|_V^2 + \lambda |a_*^k|_{\mathcal{R}}^2 \\ &\leq |v(\nabla u, a_*^k)|_V^2 + \lambda |a_*^k|_{\mathcal{R}}^2 \end{aligned}$$

(the last inequality follows from the consequence stated before Theorem 1').

To show that a_λ^k remains bounded, it is enough to bound the right-hand side of (6) by λ . Now $|a_*^k|_{\mathcal{R}} \leq |a_*|_{\mathcal{R}}$, and

$$\begin{aligned} |v(\nabla u, a_*^k)|_V &\leq |v(\nabla u, a_*^k) - v(\nabla u_*, a_*^k)|_V + |v(\nabla u_*, a_*^k) - v(\nabla u_*, a_*)|_V \\ &\leq |a_*|_\infty |\nabla u - \nabla u_*|_{L^2} + |\nabla u_*|_{L^2} |a_*^k - a_*|_{\mathcal{R}}. \end{aligned}$$

So, $|v(\nabla u, a_*^k)|_V^2$ is indeed bounded in terms of λ .

Proof of theorem. Assume otherwise that

$$\begin{aligned} \exists \varepsilon > 0 : \quad \forall n, \quad \exists \delta_n \leq 1/n, \quad \exists \lambda_n \in [l(\delta_n), r(\delta_n)], \\ \forall h_0 > 0, \quad \forall k_0 > 0, \quad \exists h_n \leq h_0, \quad \exists k_n \leq k_0 : \\ |\nabla u_n - \nabla u_*|_{L^2} \leq \delta \quad \& \quad |a_{\delta_n}^{k_n} - a_*|_{\mathcal{R}} > \varepsilon. \end{aligned}$$

Denote, for short, $a_{\lambda_n}^{k_n}$ by a_n . Taking k_0 small enough so that a_n remains bounded, we may again consider a subsequence $a_n \rightarrow \tilde{a}$ (and $a_n \rightarrow \tilde{a}$ in C^0). However,

$$v(\nabla u_n, a_n) = v^{h_n}(\nabla u_n, a_n) + v(\nabla u_n, a_n) - v^{h_n}(\nabla u_n, a_n).$$

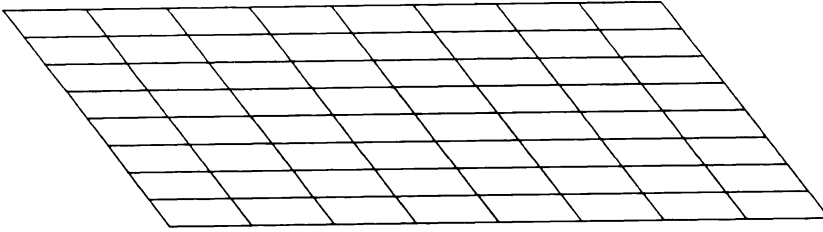
Now

$$\begin{aligned} |v^{h_n}(\nabla u_n, a_n)|^2 &\leq J_{\lambda_n}^{h_n}(\nabla u_n, a_n) \\ &\leq |v(\nabla u_n, a_n^{k_n})|_V^2 + \lambda_n |a_n^{k_n}|_{\mathcal{R}}^2 \end{aligned}$$

(see (6)). The last quantity tends to zero as $n \rightarrow \infty$ by continuity of v and $\lim \lambda_n = 0$.

TABLE 1
Example 1.

λ	No. iter.	J_e	J_1	J_2	J_λ	ϵ
0	16	$2.24 \cdot 10^{-6}$	1.13	$1.10 \cdot 10^3$	$2.24 \cdot 10^{-6}$.48
10^{-2}	202	$1.23 \cdot 10^{-9}$	$6.70 \cdot 10^{-7}$	$-3.34 \cdot 10^{-4}$	$-3.33 \cdot 10^{-6}$.000028
10^{-3}	25	$2.47 \cdot 10^{-9}$	$1.80 \cdot 10^{-7}$	$-2.03 \cdot 10^{-4}$	$-2.01 \cdot 10^{-7}$.000070
10^{-4}	12	$7.19 \cdot 10^{-8}$	$7.65 \cdot 10^{-6}$	$4.87 \cdot 10^{-3}$	$5.60 \cdot 10^{-7}$.00057
10^{-5}	14	$2.05 \cdot 10^{-7}$	$6.67 \cdot 10^{-5}$	$1.08 \cdot 10^{-1}$	$1.29 \cdot 10^{-6}$.00095

FIG. 1. Exact starting guess. $n = 8$, $\lambda = 0.0000\text{E} + 00$.

Finally,

$$\begin{aligned}
 v(\nabla u_n, a_n) - v^{h_n}(\nabla u_n, a_n) &= v(\nabla u_n, a_n) - v(\nabla u_*, a_n) \\
 &\quad + v(\nabla u_*, a_n) - v(\nabla u_*, \tilde{a}) \\
 &\quad + v(\nabla u_*, \tilde{a}) - v^{h_n}(\nabla u_*, \tilde{a}) \\
 &\quad + v^{h_n}(\nabla u_*, \tilde{a}) - v^{h_n}(\nabla u_*, a_n) \\
 &\quad + v^{h_n}(\nabla u_*, a_n) - v^{h_n}(\nabla u_n, a_n),
 \end{aligned}$$

and all these differences tend to zero; the third one by an appropriate choice of $h_0(\lambda_n)$. As before, this contradicts $\lim v(\nabla u_n, a_n) = v(\nabla u_*, \tilde{a}) \neq 0$.

3.3. Results. This method was tested on a few examples. For each example, we tried different values of the parameter λ and we show the corresponding equation error J_e , and the smoothing term broken down in J_1 and J_2 , contributions from the first- and second-order derivatives. We also keep track of the number of conjugate gradient iterations and the relative error ϵ (in the C^0 -norm) of the computed to the real a .

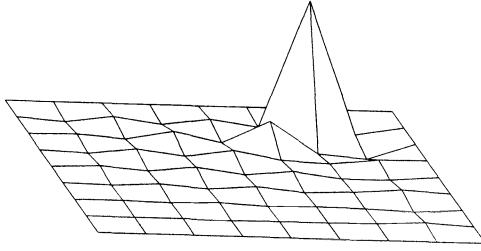
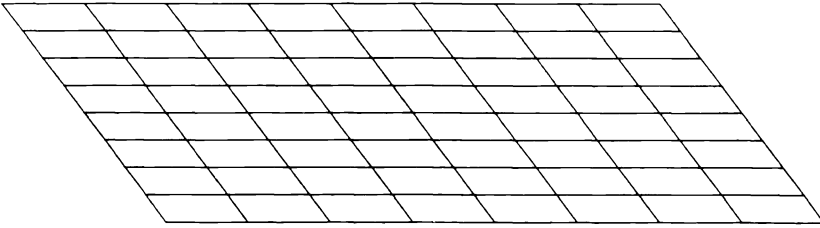
3.3.1. Example 1. In this example, let

$$\begin{aligned}
 u &= -(x + y)^2, \\
 f &\equiv 4,
 \end{aligned}$$

$$g = \begin{cases} 2(y - 1) & \text{on the left side of } \Omega, \\ 2(x - 1) & \text{at the bottom,} \\ -2(y + 1) & \text{on the right,} \\ -2(x + 1) & \text{on top,} \end{cases}$$

(so the solution is $a \equiv 1$).

We tried to see how the method behaves as the starting guess and λ vary. Figure 1 shows the result for an exact starting guess and no smoothing. Figure 2 shows that absence of smoothing may seriously affect the result if the starting guess is bad. Table 1 shows the result of runs for the same perturbed starting guess and different values of λ . Figure 3 shows the result for $\lambda = 10^{-3}$.

FIG. 2. *Perturbed starting guess. $n = 8$, $\lambda = 0.0000\text{E} + 00$.*FIG. 3. *Perturbed starting guess. $n = 8$, $\lambda = 0.1000\text{E} - 02$.*

In this example, by choice of a , $J_S(a)$ should be zero at the solution, and very small nearby. The results for $\lambda = 10^{-2}$ and 10^{-3} illustrate the numerical difficulty of giving the smoothing term too much weight:

(i) Convergence of the conjugate gradient method is very slow (202 iterations for $\lambda = 10^{-2}$);

(ii) the J_2 portion of J_S actually takes negative values.

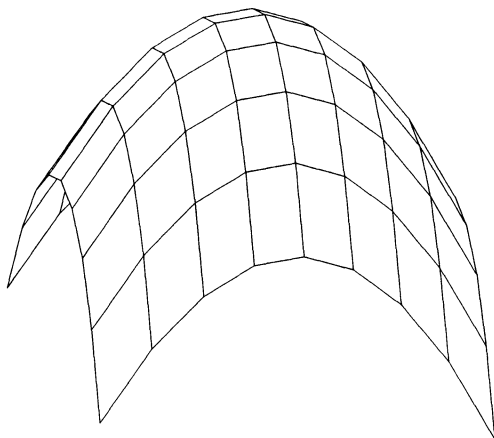
Both are symptoms of the same phenomenon: $J_2(a)$ is a quadratic form that is not positive definite ($J_2(a) = 0$ if q is constant). Its discrete version is represented by a matrix with some eigenvalues very near 0; we performed a singular-value decomposition of it (actually, of its restriction to an element patch; but the whole matrix is obtained by patching up such matrices) and we found no negative eigenvalues, but the smallest eigenvalues are small indeed ($\simeq 10^{-15}$). When a is constant (so that all the degrees of freedom corresponding to function values are nearly constant and those corresponding to derivatives nearly zero), loss of significance occurs when evaluating J_2 , caused by canceling. This behaviour disappears when λ is made small enough, for then the solution q departs from the null variety of J_2 , less importance being given to J_2 . At any rate, a negative value for J_2 shows that the solution is nearly flat, and hence is a very good approximation since, in such cases, there is no need to refine the mesh. The only trouble comes from the ill conditioning of J_2 ; this is where it is important to reduce λ .

3.3.2. Example 2. In this example, let

$$u = -(x + y),$$

$$f = -2(x + y),$$

$$g = \begin{cases} 2 - y^2 & \text{on the left side of } \Omega, \\ 2 - x^2 & \text{at the bottom,} \\ y^2 - 2 & \text{on the right,} \\ x^2 - 2 & \text{on top,} \end{cases}$$

FIG. 4. $a = 1 + (1+x)(1-x) + (1+y)(1-y)$.

The shape of the true a , $1 + (1+x)(1-x) + (1+y)(1-y)$, is shown in Fig. 4. Here again, we examined the effect of the starting guess by taking an exact one (Fig. 5), a “low” one ($a \equiv 1$ inside Ω ; see Fig. 6), and a perturbation of the latter ($p \equiv 1$ inside Ω with a spike down at a single node; see Fig. 7). This confirms what we observed with the first example that, without smoothing, variations in the starting guess tend to produce jagged and readily oscillatory solutions. With the same perturbed starting guess, we varied λ ; the results are summarised in Table 2. Figure 8 shows the solution for $\lambda = 10^{-3}$.

TABLE 2
Example 2.

λ	No. iter.	J_e	J_1	J_2	J_λ	ϵ
0	21	$5.09 \cdot 10^{-6}$	28.8	$8.50 \cdot 10^4$	$5.09 \cdot 10^{-6}$.39
10^{-2}	103	$1.66 \cdot 10^{-1}$	4.17	53.5	$7.43 \cdot 10^{-1}$.18
10^{-3}	67	$3.92 \cdot 10^{-2}$	8.12	81.6	$1.33 \cdot 10^{-1}$.067
10^{-4}	24	$2.40 \cdot 10^{-2}$	9.92	$1.28 \cdot 10^2$	$3.77 \cdot 10^{-2}$.068
10^{-5}	24	$1.15 \cdot 10^{-2}$	10.5	$6.85 \cdot 10^2$	$1.84 \cdot 10^{-2}$.070

3.2.3. Example 3. Here, we make the choice of a true a that oscillates; of course, we keep the oscillation period large with respect to the mesh, otherwise we cannot expect much.

$$\begin{aligned}
 u &= (x + y), \\
 f &= (1 - y^2) \pi \cos \pi(1 + x) - 2y \sin \pi(1 + x), \\
 g &= \begin{cases} -1 & \text{on the left side of } \Omega, \\ -1 & \text{at the bottom,} \\ 1 & \text{on the right,} \\ 1 & \text{on top.} \end{cases}
 \end{aligned}$$

Therefore, $a = 1 - (1+y)(1-y) \sin \pi(1+x)$ as shown in Fig. 9. Figure 10 shows the result of taking $a \equiv 1$ as a starting guess. A perturbed starting guess is obtained by adding a spike to the previous one; here again, the solution is sluggish in departing

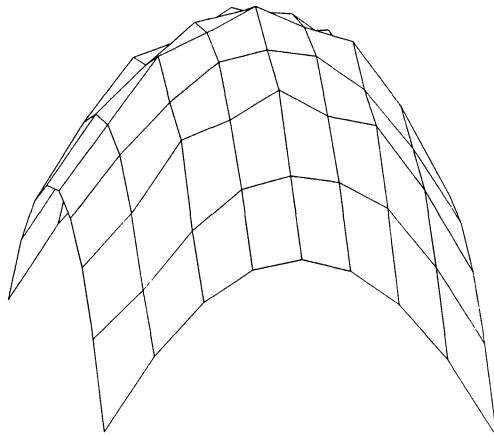


FIG. 5. *Exact starting guess.* $n = 8$, $\lambda = 0.0000\text{E} + 00$.

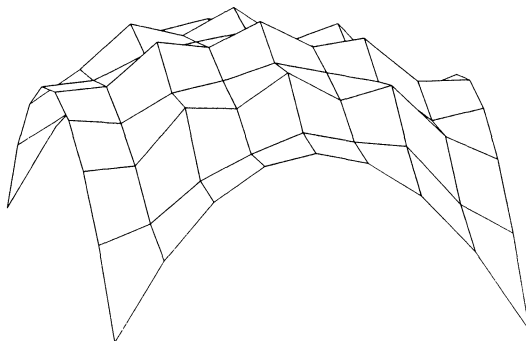


FIG. 6. *Low starting guess.* $n = 8$, $\lambda = 0.0000\text{E} + 00$.

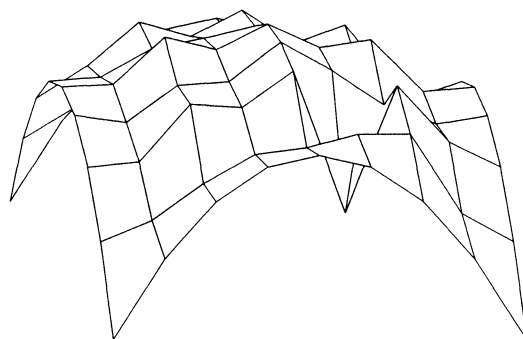
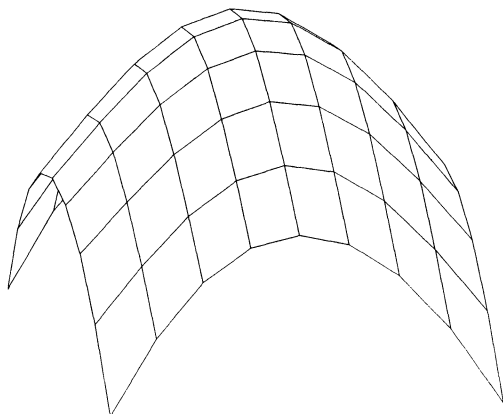


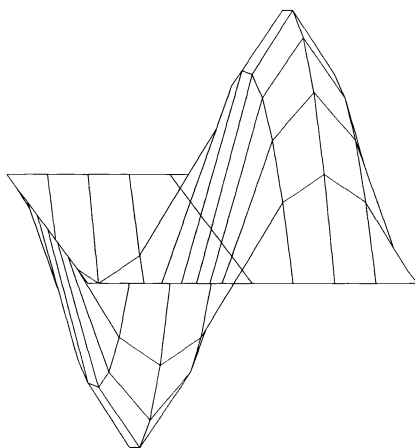
FIG. 7. *Perturbed starting guess.* $n = 8$, $\lambda = 0.0000\text{E} + 00$.

from the starting guess without smoothing (Fig. 11). For the perturbed starting guess, test runs are summarised in Table 3. Figure 12 shows the results for $\lambda = 10^{-6}$.

4. Identification of discontinuous transmissivity. We now turn to the situation where a is known to have jump discontinuities across the domain. This will be the case if the material making up the domain is not homogeneous at the macroscopic

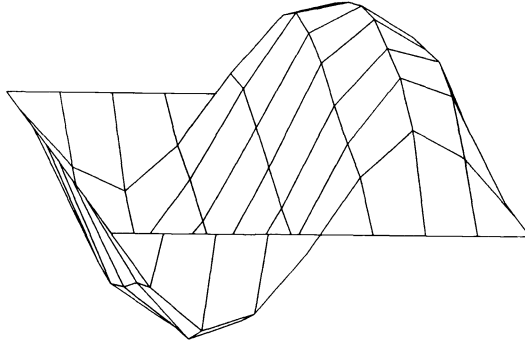
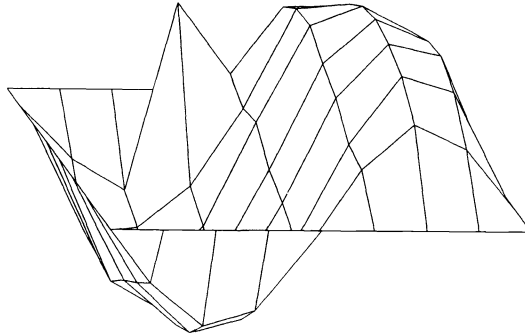
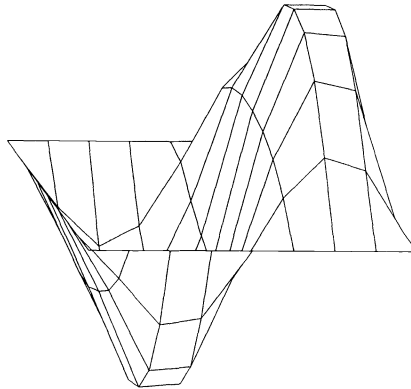
FIG. 8. *Perturbed starting guess.* $n = 8$, $\lambda = 0.1000\text{E} - 02$.TABLE 3
Example 3.

λ	No. iter.	J_e	J_1	J_2	J_λ	ϵ
0	13	$7.52 \cdot 10^{-6}$	10.4	$1.73 \cdot 10^4$	$7.52 \cdot 10^{-6}$.34
10^{-2}	96	$2.74 \cdot 10^{-1}$.00435	1.69	$2.91 \cdot 10^{-1}$.49
10^{-3}	49	$1.30 \cdot 10^{-1}$	1.53	64.5	$1.96 \cdot 10^{-1}$.36
10^{-4}	23	$2.60 \cdot 10^{-2}$	7.67	$3.53 \cdot 10^2$	$6.21 \cdot 10^{-2}$.19
10^{-5}	14	$8.93 \cdot 10^{-3}$	10.9	$9.28 \cdot 10^2$	$1.83 \cdot 10^{-2}$.17
10^{-6}	20	$6.75 \cdot 10^{-4}$	12.2	$3.32 \cdot 10^3$	$4.01 \cdot 10^{-3}$.16
10^{-7}	20	$1.59 \cdot 10^{-5}$	12.5	$4.85 \cdot 10^3$	$5.02 \cdot 10^{-4}$.18
10^{-8}	13	$8.64 \cdot 10^{-6}$	10.1	$1.61 \cdot 10^4$	$1.70 \cdot 10^{-4}$.31

FIG. 9. $a = 1 - (1 + y)(1 - y) \sin(\pi(1 + x))$.

scale; in our model, this translates into fracture curves.

To simplify the analysis, we assume that there is a single such fracture curve across Ω , and that it is smooth. It divides Ω into two subdomains, Ω_1 and Ω_2 . Our assumptions on the data are the same as before: we know f , ∇u , and the flux $a(\partial u)/(\partial n)$ at the boundary of Ω . We also assume conservation of flux across the fracture curve;

FIG. 10. *Flat starting guess. $n = 8$, $\lambda = 0.0000\text{E} + 00$.*FIG. 11. *Perturbed starting guess. $n = 8$, $\lambda = 0.0000\text{E} + 00$.*FIG. 12. *Perturbed starting guess. $n = 8$, $\lambda = 0.1000\text{E} - 05$.*

this is the *Rankine–Hugoniot condition*: if a_1 and u_1 are the transmissivity and potential in Ω_1 , and a_2 and u_2 the corresponding quantities in Ω_2 , then along the fracture curve

$$(7) \quad a_1 \frac{\partial u_1}{\partial n} + a_2 \frac{\partial u_2}{\partial n} = 0,$$

where $\partial u_i / \partial n$ is the derivative of u in the outer normal direction to Ω_i .

Assuming that we have no data on the flux along the fracture curve, we suggest extending the equation-error method to this case. The class of admissible a 's consists now of pairs (a_1, a_2) such that:

- (i) Ω_1, Ω_2 form a partition of Ω , separated by a smooth curve C ;
- (ii) $a_1 \in H^2(\Omega_1)$ and $a_2 \in H^2(\Omega_2)$;
- (iii) a_1 and a_2 (along with $\nabla u_1, \nabla u_2$) satisfy condition (7) along C .

In this setup, (3) is still a valid formulation; we may again define the equation error $v(\nabla u, a)$ by (4), and, in the same spirit as before, we want to minimise a suitable modification of $J_e(\nabla u, a)$. The procedure is then as follows: For each admissible smooth curve C satisfying condition (i), consider the set of a 's satisfying (ii) and (iii). The problem of minimising (C still fixed)

$$J_\lambda(a) := J_e(\nabla u, a) + \lambda(|a_1|_{H^2(\Omega_1)}^2 + |a_2|_{H^2(\Omega_2)}^2)$$

over all admissible a 's is stable; we call its solution a_C . Then minimise $J_\lambda(a_C)$ over all admissible curves C .

4.1. Numerical implementation. In this setup, we assume that C joins the midpoints of the top and bottom of Ω . While this is a simplistic assumption (in practice, we do not know where the lines of discontinuity are), it may be justified by the fact that discontinuities in the boundary flux may be "seen." We approximate it by a broken line; the mesh is generated by horizontal scaling on each side of C . Since the mesh is no longer regular, the basic "pseudo-tensor" splines obtained by patching up contiguous shapes have discontinuous derivatives across horizontal lines, so that now A_k is an external approximation of $H^2(\Omega_i)$ (see [Te]). If a_C minimises $J_C(a)$ given C , the optimum function is

$$C \longmapsto \Phi(C) := J_C(a_C).$$

The problem now is to minimise $\Phi(C)$, where C is constrained to stay in the domain; the inner loop computes $\Phi(C)$, the outer loop moves C . There are library numerical routines that perform such constrained minimisation problems (for example, in the NAG library), assuming that Φ is nearly differentiable in its vector argument (but not requiring the user to specify the gradient) and using quasi-Newton steps. A typical solution would require, however quite a few evaluations of Φ and it would not have been advisable to do this on the VAX/780 which we used because each single evaluation of Φ , using a tiling of Ω with 16 mesh nodes, would take a few hours if the conjugate gradient method was carried to term (even with a tolerance on a^h of, say, 10^{-3}).

Since convergence of the conjugate gradient search was found to be fast at first, then considerably slow, we truncated the number of iterations to less than 10, and had to trust the results thus obtained. As for the outer loop, we generated a few profiles C^h in the neighbourhood of the true one, and recorded the value $\Phi(C^h)$ yielded by the conjugate gradient search for each C^h . The purpose of this simulation of the constrained problem corresponding to the outer loop is to check whether the true profile indeed gives the lowest value for Φ . A few experiments tended to justify this rudimentary procedure, in that for different samples C^h , we truncated the conjugate gradient procedure at different stages (10 versus 20 iterations), and we observed that at each stage the ordering of the values $\Phi(C^h)$ was roughly the same; also, that the shape of the solution (at least those values corresponding to values of a itself, not its derivatives) settled rather fast.

4.2. Results. We tested the method on two examples; one where C is a straight line, and the other where C is an arc of parabola. In each case, we generated six curves representing perturbations of C and three randomly generated curves, so that we had ten curves including C itself. With different values of λ we recorded, for each curve (or profile), the quantities J_e , J_2 , and J_λ , as well as the relative error ϵ . The mesh number in both examples is $n = 4$.

4.2.1. Example 4. In this example, C is the line $x = 0$, dividing Ω in Ω_1 , and Ω_2 on the left and right.

$$\begin{aligned} u_1 &= 2x + y, \\ u_2 &= x + y, \\ f &= 0, \\ g_1 &= \begin{cases} 1 & \text{on top,} \\ -2 & \text{on the left,} \\ -1 & \text{on the bottom,} \end{cases} \\ g_2 &= \begin{cases} 2 & \text{on top,} \\ 2 & \text{on the right,} \\ -2 & \text{on the bottom.} \end{cases} \end{aligned}$$

Profiles are represented by their intercepts x_1, x_2, x_3 on the horizontal lines of the grid. We denote them by $\pm i$ ($1 \leq i \leq 3$) and a, b, c ; the first are perturbations of C , and the last three are generated from a normal distribution centered about C :

$$\begin{aligned} a_{-1} &\leftrightarrow (-.1, 0, 0), \\ a_1 &\leftrightarrow (.1, 0, 0), \\ a_{-2} &\leftrightarrow (0, -.1, 0), \\ a_2 &\leftrightarrow (0, .1, 0), \\ a_{-3} &\leftrightarrow (0, 0, -.1), \\ a_3 &\leftrightarrow (0, 0, .1), \\ a_0 &\leftrightarrow (0, 0, 0), \\ a_a &\leftrightarrow (.11, .26, .32), \\ a_b &\leftrightarrow (-.01, .3, -.09), \\ a_c &\leftrightarrow (-.15, .05, .04). \end{aligned}$$

Tables 4 and 5 show the results for $\lambda = 10^{-5}$ and 10^{-4} . In both cases, the method picks the profile C itself; Table 6 shows the results for the true profile and various values of λ . As could be expected for this example, high values of λ give better results.

Figure 13 shows the computed solution for $\lambda = 10^{-3}$, and Fig. 14 shows the computed solution in the absence of smoothing ($\lambda = 0$).

4.2.2. Example 5. In this example, C is the curve of equation $4x - y^2 + 1 = 0$

TABLE 4
Example 4, $\lambda = 10^{-5}$.

Profile	J_e	J_2	J_λ	ϵ
-1	$1.13 \cdot 10^{-3}$	14.7	$1.29 \cdot 10^{-3}$	$.91 \cdot 10^{-2}$
1	$1.03 \cdot 10^{-3}$	13.6	$1.18 \cdot 10^{-3}$	$.10 \cdot 10^{-1}$
-2	$6.29 \cdot 10^{-4}$	13.8	$7.82 \cdot 10^{-4}$	$.11 \cdot 10^{-2}$
2	$9.07 \cdot 10^{-4}$	12.6	$1.05 \cdot 10^{-3}$	$.11 \cdot 10^{-1}$
-3	$9.91 \cdot 10^{-4}$	13.7	$1.14 \cdot 10^{-3}$	$.10 \cdot 10^{-1}$
3	$1.44 \cdot 10^{-3}$	13.8	$1.59 \cdot 10^{-3}$	$.98 \cdot 10^{-2}$
0	$1.21 \cdot 10^{-4}$	12.0	$2.53 \cdot 10^{-4}$	$.73 \cdot 10^{-3}$
<i>a</i>	$7.57 \cdot 10^{-4}$	47.0	$1.27 \cdot 10^{-3}$	$.91 \cdot 10^{-2}$
<i>b</i>	$1.24 \cdot 10^{-3}$	50.0	$1.77 \cdot 10^{-3}$	$.24 \cdot 10^{-1}$
<i>c</i>	$1.20 \cdot 10^{-3}$	48.5	$1.74 \cdot 10^{-3}$	$.23 \cdot 10^{-1}$

TABLE 5
Example 4, $\lambda = 10^{-4}$.

Profile	J_e	J_2	J_λ	ϵ
-1	$1.09 \cdot 10^{-3}$	6.32	$1.74 \cdot 10^{-3}$	$.10 \cdot 10^{-1}$
1	$1.05 \cdot 10^{-3}$	6.42	$1.72 \cdot 10^{-3}$	$.11 \cdot 10^{-1}$
-2	$6.46 \cdot 10^{-4}$	5.74	$1.24 \cdot 10^{-3}$	$.11 \cdot 10^{-1}$
2	$1.22 \cdot 10^{-1}$	14.0	$1.22 \cdot 10^{-1}$	$.54 \cdot 10^{-3}$
-3	$1.06 \cdot 10^{-3}$	5.49	$1.63 \cdot 10^{-3}$	$.11 \cdot 10^{-1}$
3	$1.30 \cdot 10^{-3}$	6.44	$1.97 \cdot 10^{-3}$	$.10 \cdot 10^{-1}$
0	$7.39 \cdot 10^{-5}$	4.82	$5.73 \cdot 10^{-4}$	$.40 \cdot 10^{-3}$
<i>a</i>	$1.12 \cdot 10^{-3}$	18.2	$3.06 \cdot 10^{-3}$	$.11 \cdot 10^{-1}$
<i>b</i>	$1.39 \cdot 10^{-3}$	23.5	$3.84 \cdot 10^{-3}$	$.28 \cdot 10^{-1}$
<i>c</i>	$1.60 \cdot 10^{-3}$	23.8	$4.09 \cdot 10^{-3}$	$.25 \cdot 10^{-1}$

TABLE 6
Example 4, true fracture line.

λ_s	J_e	J_2	J_λ	ϵ
0	$1.30 \cdot 10^{-4}$	13.5	$1.41 \cdot 10^{-4}$	$.78 \cdot 10^{-3}$
10^{-3}	$1.18 \cdot 10^{-4}$	0.579	$7.18 \cdot 10^{-4}$	$.94 \cdot 10^{-4}$
10^{-4}	$7.39 \cdot 10^{-5}$	4.82	$5.73 \cdot 10^{-4}$	$.40 \cdot 10^{-3}$
10^{-5}	$1.21 \cdot 10^{-4}$	12.0	$2.53 \cdot 10^{-4}$	$.73 \cdot 10^{-3}$
10^{-6}	$1.29 \cdot 10^{-4}$	13.4	$1.54 \cdot 10^{-4}$	$.77 \cdot 10^{-3}$
10^{-7}	$1.30 \cdot 10^{-4}$	13.5	$1.43 \cdot 10^{-4}$	$.78 \cdot 10^{-3}$

as shown in Fig. 15.

$$u_1 = 4x - y^2 + 1 + x,$$

$$u_2 = 2x - y^2/2 + 1/2 + x,$$

$$f_1 = 6 + 3y^2,$$

$$f_2 = 5 + 3y^2,$$

$$g_1 = \begin{cases} -7 & \text{on top,} \\ -5(3 + y^2/2) & \text{on the left,} \\ -7 & \text{on the bottom,} \end{cases}$$

$$g_2 = \begin{cases} -6 & \text{on top,} \\ 3(5 + y^2) & \text{on the right,} \\ -6 & \text{at the bottom.} \end{cases}$$

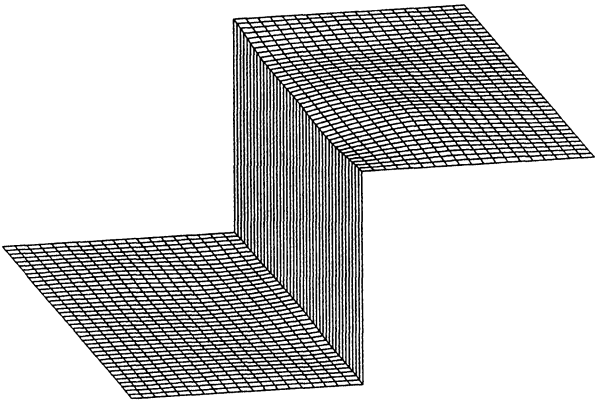


FIG. 13. $x_1 = 0.0000$, $x_2 = 0.000$, $x_3 = 0.0000$.

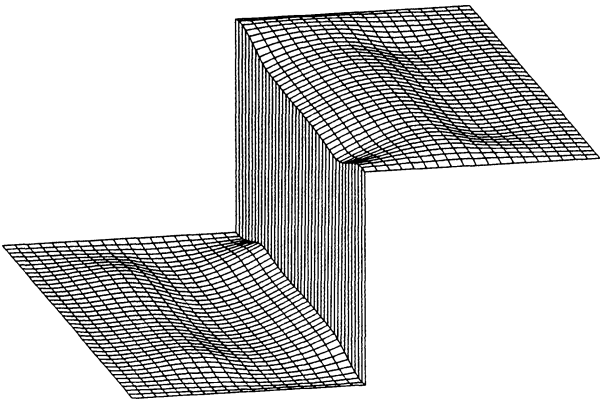


FIG. 14. $x_1 = 0.0000$, $x_2 = 0.0000$, $x_3 = 0.0000$.

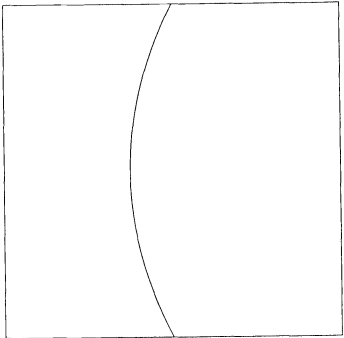


FIG. 15. *Fracture line.*

The solution is then $a_1 = 3 + y^2/2$, $a_2 = 5 + y^2$ as shown in Fig. 16.

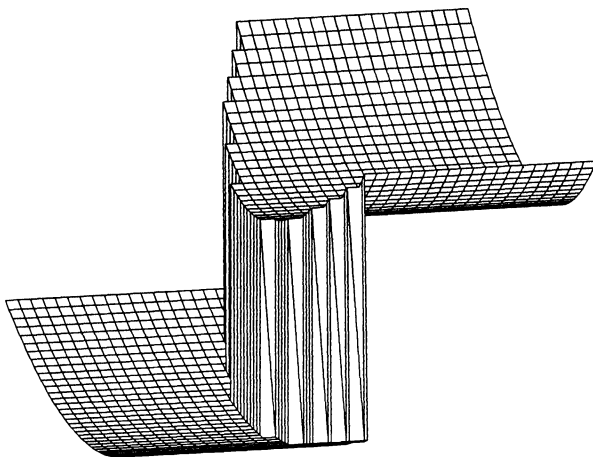


FIG. 16. $a = 3 + y^2/2$ at left, $5 + y^2$ at right.

The profiles now are

$$\begin{aligned}
 a_{-1} &\leftrightarrow (-.2875, -.25, -.1875), \\
 a_1 &\leftrightarrow (-.0875, -.25, -.1875), \\
 a_{-2} &\leftrightarrow (-.1875, -.35, -.1875), \\
 a_2 &\leftrightarrow (-.1875, -.15, -.1875), \\
 a_{-3} &\leftrightarrow (-.1875, -.25, -.2875), \\
 a_3 &\leftrightarrow (-.1875, -.25, -.0875), \\
 a_0 &\leftrightarrow (-.1875, -.25, -.1875), \\
 a_a &\leftrightarrow (-.08, -.25, -.34), \\
 a_b &\leftrightarrow (.07, .05, -.14), \\
 a_c &\leftrightarrow (.13, -.33, -.15).
 \end{aligned}$$

Tables 7 and 8 show the results for $\lambda = 10^{-5}$ and 10^{-6} , respectively. For the first value of λ , we had to pass up the outlier profile a_{-3} which was run early on, for unlike the other profiles tested, we allowed the conjugate gradient procedure for this one to run for more than 20 iterations. So the method picks profile a_{-2} , which is indeed a better approximation of C than a_0 , which is an inner envelope lying on one side of C .

For comparison, Table 9 is for $\lambda = 0$.

Figures 17 and 18 show the solution for $\lambda = 10^{-6}$ and profiles -3 and 3 , respectively. Figure 19 shows the solution for the same profile in the absence of smoothing.

Note. A similar approach to the one developed here has been implemented by Patricia Lamm in the case of the observation error criterion; see [La]. In her paper, a is assumed to be piecewise constant, and compactification in a is enforced by bound constraints instead of smoothing.

5. Conclusion. The transmissivity identification problem is ill posed, and no solution method is entirely satisfactory. The method under consideration, compared with the observation error method, presents the advantage of easier analysis stemming from convexity of the objective function. Some results, such as the validity of the Tikhonov–Arsenin parameter selection strategy, carry through with some of their hypotheses (usually impossible to check in the case of the observation-error method)

TABLE 7
Example 5, $\lambda = 10^{-5}$.

Profile	J_e	J_2	J_λ	ϵ
-1	$1.84 \cdot 10^{-2}$	$4.86 \cdot 10^2$	$2.37 \cdot 10^{-2}$	$.57 \cdot 10^{-2}$
1	$2.15 \cdot 10^{-2}$	$4.73 \cdot 10^2$	$2.72 \cdot 10^{-2}$	$.69 \cdot 10^{-2}$
-2	$6.78 \cdot 10^{-3}$	$4.98 \cdot 10^2$	$1.22 \cdot 10^{-2}$	$.58 \cdot 10^{-2}$
2	$1.12 \cdot 10^{-2}$	$6.03 \cdot 10^2$	$1.80 \cdot 10^{-2}$	$.16 \cdot 10^{-2}$
-3*	$1.30 \cdot 10^{-3}$	$5.80 \cdot 10^2$	$7.77 \cdot 10^{-3}$	$.75 \cdot 10^{-2}$
3	$2.15 \cdot 10^{-2}$	$4.73 \cdot 10^2$	$2.72 \cdot 10^{-2}$	$.69 \cdot 10^{-2}$
0	$1.42 \cdot 10^{-2}$	$4.82 \cdot 10^2$	$1.95 \cdot 10^{-2}$	$.75 \cdot 10^{-3}$
a	$4.30 \cdot 10^{-2}$	$5.05 \cdot 10^2$	$4.93 \cdot 10^{-2}$	$.92 \cdot 10^{-2}$
b	$2.90 \cdot 10^{-2}$	$5.52 \cdot 10^2$	$3.55 \cdot 10^{-2}$	$.62 \cdot 10^{-3}$
c	$4.16 \cdot 10^{-2}$	$6.32 \cdot 10^2$	$4.86 \cdot 10^{-2}$	$.13 \cdot 10^{-1}$

TABLE 8
Example 5, $\lambda = 10^{-6}$.

Profile	J_e	J_2	J_λ	ϵ
-1	$1.24 \cdot 10^{-2}$	$4.94 \cdot 10^2$	$1.93 \cdot 10^{-2}$	$.57 \cdot 10^{-2}$
1	$4.14 \cdot 10^{-2}$	$6.44 \cdot 10^2$	$4.26 \cdot 10^{-2}$	$.13 \cdot 10^{-1}$
-2	$7.06 \cdot 10^{-3}$	$5.19 \cdot 10^2$	$7.99 \cdot 10^{-3}$	$.58 \cdot 10^{-2}$
2	$1.15 \cdot 10^{-2}$	$6.14 \cdot 10^2$	$1.28 \cdot 10^{-2}$	$.16 \cdot 10^{-2}$
-3	$1.84 \cdot 10^{-2}$	$4.94 \cdot 10^2$	$1.93 \cdot 10^{-2}$	$.57 \cdot 10^{-2}$
3	$2.15 \cdot 10^{-2}$	$4.81 \cdot 10^2$	$2.29 \cdot 10^{-2}$	$.69 \cdot 10^{-2}$
0	$1.44 \cdot 10^{-2}$	$4.94 \cdot 10^2$	$1.53 \cdot 10^{-2}$	$.77 \cdot 10^{-3}$
a	$4.29 \cdot 10^{-2}$	$5.16 \cdot 10^2$	$4.46 \cdot 10^{-2}$	$.92 \cdot 10^{-2}$
b	$2.92 \cdot 10^{-2}$	$5.65 \cdot 10^2$	$3.06 \cdot 10^{-2}$	$.62 \cdot 10^{-2}$
c	$4.14 \cdot 10^{-2}$	$6.44 \cdot 10^2$	$4.26 \cdot 10^{-2}$	$.13 \cdot 10^{-1}$

TABLE 9
Example 5, $\lambda = 0$.

Profile	J_e	J_2	J_λ	ϵ
-2	$7.01 \cdot 10^{-3}$	$5.21 \cdot 10^2$	$7.42 \cdot 10^{-3}$	$.58 \cdot 10^{-2}$

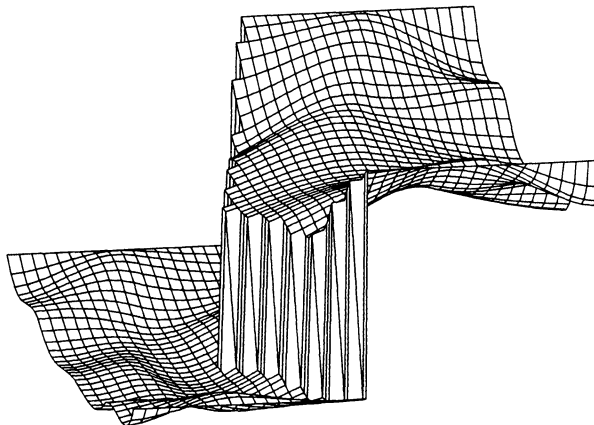
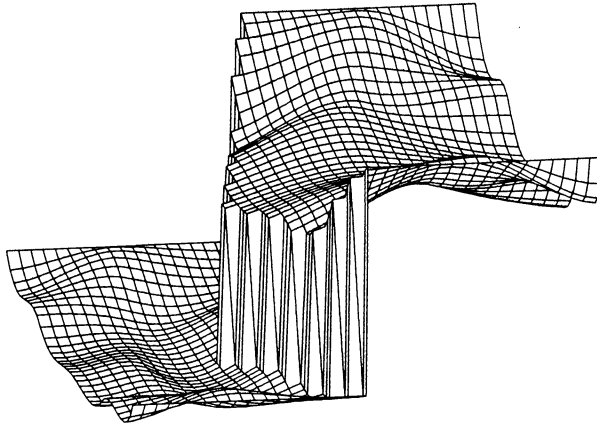
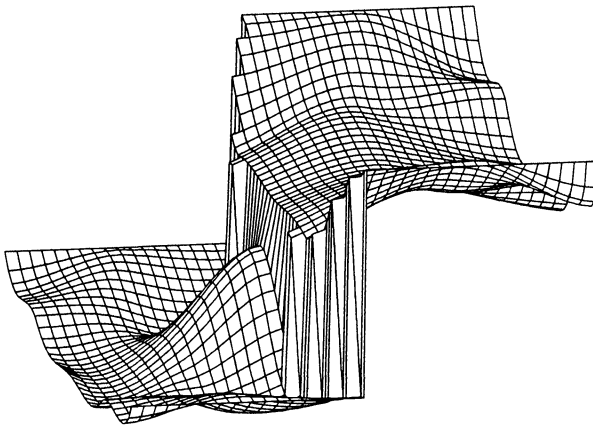


FIG. 17. $x_1 = -0.1875$, $x_2 = -0.3500$, $x_3 = -0.1875$.

automatically satisfied. This indicates that the equation error may be a more natu-

FIG. 18. $x_1 = -0.1875$, $x_2 = -0.1500$, $x_3 = -0.1875$.FIG. 19. $x_1 = -0.1875$, $x_2 = -0.3500$, $x_3 = -0.1875$.

ral criterion than the observation error: it reflects less the direct readings of u , and depends more on the coupling between a and u . Both methods require the use of some smoothing for stability. The main disadvantage of this method is the increased complexity in computation; computing J_2 , the second-order portion of the H^2 -norm, is the main culprit for the CPU time used (40 seconds per function evaluation on the VAX/780). The same term accounts also for defective convergence of the conjugate gradient method, as we noted when we unduly increased λ . On the other hand, while the theoretical convergence results rely on H^2 -smoothing, we did observe the remarkable fact that overspecifying the boundary data (as in §4) seems enough to enforce smoothing on the computed a (without the help of H^2 -smoothing).

Acknowledgments. The author is grateful to David Russell for proposing the topic for this work and for his unfaltering encouragement, patience and support; to Seymour Parter and Grace Wahba for their guidance and many useful discussions; and, to Fritz Colonius who told him about Tikhonov regularisation. He also thanks an anonymous referee for pointing out references [A], [CG], [HS], and [KL].

REFERENCES

- [A] G. ALESSANDRINI, *An identification problem for an elliptic equation in two variables*, Ann. Mat. Pura Appl., 145 (1986), pp. 265–296.
- [BK1] H. T. BANKS AND K. KUNISCH, *Estimation techniques for distributed parameter systems*, Birkhäuser, Boston, 1989.
- [BK2] R. BELLMAN AND R. KALABA, *Quasilinearization and Nonlinear Boundary Value Problems*, Elsevier, New York, 1965.
- [BW] K.-J. BATHE AND E. L. WILSON, *Numerical Methods in Finite Element Analysis*, Prentice-Hall, New York, 1976.
- [CY] S. CHANG AND W. W.-G. YEH, *A proposed algorithm for the solution of the large-scale inverse problem in groundwater*, Water Resour. Res., 12 (1976), pp. 365–374.
- [C] G. CHAVENT, *About the stability of the optimal control solution of inverse problems*, in Inverse and Improperly Posed Problems in Differential Equations, G. Anger, ed., Akademie-Verlag, Berlin, 1979, pp. 74–86.
- [CDL] G. CHAVENT, M. DUPUY, AND P. LEMONNIER, *History matching by use of optimal theory*, Soc. Petrol. Engr. J., 15 (1975), pp. 74–86.
- [CS] W. H. CHEN AND J. H. SEINFELD, *Estimation of spatially varying parameters in partial differential equations*, Internat. J. Control, 15 (1972), pp. 487–495.
- [CG] C. CHICONE AND J. GERLACH, *A note on the identifiability of distributed parameters in elliptic equations*, SIAM J. Math. Anal., 18 (1987), pp. 1378–1384.
- [EdM] Y. EMSELLEM AND G. DE MARSILY, *An automatic solution for the inverse problem*, Water Resour. Res., 7 (1971), pp. 1264–1283.
- [F] R. S. FALK, *Error estimates for the numerical identification of a variable coefficient*, Math. Comput., 140 (1983), pp. 537–546.
- [FP] E. O. FRIND AND G. F. PINDER, *Galerkin solution of the inverse problem for aquifer transmissivity*, Water Resour. Res., 9 (1973), pp. 1397–1410.
- [HS] K. H. HOFFMAN AND J. SPREKELS, *On the identification of coefficients of elliptic problems by asymptotic regularization*, Numer. Funct. Anal. Optim., 7 (1984–1985), pp. 157–177.
- [JJ] P. JACQUARD AND C. JAIN, *Permeability distribution from field pressure data*, Soc. Petrol. Eng. J., 5 (1965), pp. 281–294.
- [J] H. O. JAHNS, *A rapid method for obtaining a two-dimensional reservoir description from well pressure response data*, Soc. Petrol. Eng. J., 6 (1966), pp. 315–327.
- [KN] S. KITAMURA AND S. NAKAGIRI, *Identifiability of spatially-varying parameters in distributed systems of parabolic type*, SIAM J. Cont. Optim., 15 (1977), pp. 785–802.
- [Kl] D. KLEINECKE, *Use of linear programming for estimating geohydrologic parameters of groundwater basins*, Water Resour. Res., 7 (1971), pp. 367–374.
- [KL] R. V. KOHN AND B. D. LOWE, *A variational method for parameter identification*, RAIRO Modél. Mat. Anal. Numer., 22 (1988), pp. 119–158.
- [KS] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed parameter systems by regularization*, SIAM J. Cont. Optim., 23 (1985), pp. 217–241.
- [Kr] W. D. KRUGER, *Determining areal permeability distribution by calculation*, J. Petrol. Technol., (1961), pp. 691–698.
- [La] P. K. LAMM, *Isoparametric finite element methods to estimate discontinuous coefficients in two-dimensional elliptic equations*, preprint.
- [LY] A. C. LIN AND W. W.-G. YEH, *Identification of parameters in an inhomogeneous aquifer by use of the maximum principle of control and quasi-linearization*, Water Resour. Res., 10 (1974), pp. 829–838.
- [L] J.-L. LIONS, *Some aspects of modeling problems in distributed parameter systems*, in Proc. IFIP Working Conference, Rome, 1976, A. Ruberti, ed., Lecture Notes in Control and Information Sciences 1, Springer-Verlag, Berlin, 1978, pp. 11–41.
- [Mi] K. MILLER, *Least-square methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.
- [Mo] V. A. MOROZOV, *Methods for solving incorrectly posed problems*, Springer-Verlag, New York, 1984.
- [Ne] R. W. NELSON, *In-place determination of permeability distribution for heterogeneous porous media through analysis of energy dissipation*, Soc. Petrol. Eng. J., 8 (1968), pp. 33–42.
- [Nu] D. A. NUTBROWN, *Identification of parameters in a linear equation of groundwater flow*, Water Resour. Res., 11 (1975), pp. 581–588.
- [P] G. A. PHILLIPSON, *Identification of Distributed Systems*, Elsevier, New York, 1971.

- [R1] G. R. RICHTER, *Numerical identification of a spatially varying diffusion coefficient*, Math. Comp., 36 (1981), pp. 375–386.
- [R2] ———, *An inverse problem for the steady-state diffusion equation*, SIAM J. Appl. Math., 41 (1981), pp. 210–221.
- [SYD] B. SAGAR, S. YAKOWITZ, AND L. DUCKSTEIN, *A direct method for the identification of the parameters of dynamic nonhomogeneous aquifers*, Water Resour. Res., 11 (1975), pp. 563–570.
- [S] R. W. STALLMAN, *Numerical analysis of regional water levels to define aquifer hydrology*, Eos Trans. AGU, 37 (1956), pp. 451–460.
- [Té] R. TEMAM, *Numerical Analysis*, Reidel, Dordrecht, the Netherlands, 1970.
- [Ti1] A. N. TIKHONOV, *Solution of ill-posed problems and the regularization method*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 501–504; Soviet Math. Dokl., 4 (1963), pp. 1035–1038.
- [Ti2] ———, *Regularisation of ill-posed problems*, Dokl. Akad. Nauk SSSR, 153 (1963), pp. 49–52; Soviet Math. Dokl., 4 (1963), pp. 1624–1627.
- [TA] A. N. TIKHONOV AND V. Y. ARSENIN, *Solution of Ill-Posed Problems*, Winston-Wiley, New York, 1977.
- [YT] W. W.-G. YEH AND G. W. TAUXE, *Optimal identification of aquifer diffusivity using quasi-linearization*, Water Resour. Res., 7 (1971), pp. 955–962.

LINEAR SYSTEMS WITH SIGN-OBSERVATIONS*

RENÉE KOPLON[†] AND EDUARDO D. SONTAG[‡]

Abstract. This paper deals with systems that are obtained from linear time-invariant continuous- or discrete-time devices followed by a function that just provides the sign of each output. Such systems appear naturally in the study of quantized observations as well as in signal processing and neural network theory. Results are given on observability, minimal realizations, and other system-theoretic concepts. Certain major differences exist with the linear case, and other results generalize in a surprisingly straightforward manner.

Key words. observability, minimal realization, neural networks, quantization effects

AMS subject classifications. 93B07, 93B10, 93B15

1. Introduction. A central issue in current control theory and signal processing concerns the interface between, on the one hand, the continuous, physical, world and, on the other hand, discrete devices such as digital computers, capable of symbolic processing. Classical control techniques, especially for linear systems, have proved spectacularly successful in automatically regulating relatively simple systems. However, for large-scale problems, controllers resulting from the application of the well-developed theory are used as building blocks of more complex systems. The integration of these systems is often accomplished by means of ad hoc techniques that combine pattern recognition devices, various types of switching controllers, and humans—or, more recently, expert systems—in supervisory capabilities.

Recently, there has been renewed interest in the formulation of mathematical models in which this interface between the continuous and the symbolic is naturally accomplished and system-theoretic questions can be formulated and resolved for the resulting models. Successful approaches will eventually allow the interplay of modern control theory with automata theory and other techniques from computer science. This interest has motivated much research into areas such as discrete-event systems, supervisory control, and, more generally, “intelligent control systems.”

One possible first step in a systematic attack of this problem is the study of partial (discrete) measurements on the state of a continuous dynamical system. When no controls are present, this is closely related to classical work on symbolic dynamics, and in fact has been pursued in the control theory literature, where Ramadge studied in [9] the dynamical behavior of observation sequences corresponding to such systems.

If inputs are available, one of the first questions that we may address in this context is that of the nature of the information that can be deduced by a symbolic “supervisor” from data transmitted by such a “lower level” continuous device, using appropriate controls to obtain more information about the system. Here the work of Delchamps, especially in [3]–[5], is especially relevant. His work dealt with what we may call “single-experiment observability” of *constrained-output systems*, systems for which the dynamics are linear but the outputs reflect various limitations of measuring

* Received by the editors June 17, 1991; accepted for publication (in revised form) September 8, 1992. This research was supported in part by Air Force Office of Scientific Research grant AFOSR-91-0343.

[†] Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903 (koplon@hilbert.rutgers.edu).

[‡] Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903 (sontag@hilbert.rutgers.edu).

devices. These are systems, in discrete or continuous time, whose equations can be expressed as

$$(1) \quad \begin{aligned} x(t+1) \text{ [or } \dot{x}(t)] &= Ax(t) + Bu(t), \\ y(t) &= \sigma(Cx), \end{aligned}$$

for some $n \times n$ real matrix A , $n \times m$ matrix B , and $p \times n$ matrix C , and where σ is a memory-free map: $\mathbb{R}^p \rightarrow \mathbb{R}^p$ —in the case of Delchamp's work, a quantizer. (The simplest example of a constrained-output system occurs if σ is the identity. Then we are dealing with the class of all finite-dimensional linear systems. See [12] for precise definitions of "system" and related terms.) Models of the form (1) with quantizer σ arise also in a variety of other areas besides control. For instance, in signal processing, when modeling linear channels transmitting digital data from a quantized source, the channel equalization problem becomes one of systems inversion for such systems; see [2] and also the related paper [8].

In contrast to Delchamp's work, in this paper we look at the more standard notion of multiple-experiment observability, which is different for nonlinear systems from the single-experiment concept (for purely linear systems, both concepts do coincide, of course). We will be especially interested in the case in which σ simply takes the sign of each coordinate, that is, *sign-linear systems*, those for which

$$\sigma(x) = \text{sign}(x)$$

(applied to each coordinate independently), where

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Sign-linear systems correspond to the 1-bit quantization case of Delchamps' model and are also motivated by pattern recognition applications (see below), but many technical results will be given in the paper in somewhat more generality.

Among the most popular techniques in pattern classification are those based on the use of *perceptrons* or linear discriminants (see, e.g., [6], [13]). Mathematically, these are simply functions of the type

$$\mathbb{R}^n \mapsto \mathbb{R}^p, \quad v \mapsto y = \text{sign}(Cv),$$

typically with large n and small p ; again, the sign is understood as being taken in each coordinate separately. Perceptrons are used to classify input patterns $v = (v_1, \dots, v_n)$ into classes, and they form the basis of many statistical techniques. In many practical situations arising in speech processing or learning finite automata and languages (see, e.g., [7]), the vector v really represents a finite window

$$(2) \quad u(t-1), \dots, u(t-s)$$

of a sequence of m -dimensional inputs $u(1), u(2), \dots$, where the components of (2) have been listed as v (and $sm = n$). In that case, the perceptron can be understood as a sign-linear system of dimension n , with a shift-register used to store the previous inputs (2). Borrowing from the signal processing terminology, perceptrons are "finite impulse response" sign-linear systems. As such, they are not suited to modeling time dependencies and recurrences in the data. More general sign-linear systems are called

for, and this motivated the introduction of such systems in [1], using the name “infinite impulse response” again by analogy to the classical linear case. In that paper, the authors studied practical problems of systems identification but did not address the more system-theoretic types of questions with which this paper deals.

As a final reason for studying sign-linear systems, we point out that such systems provide a natural class of nonsmooth nonlinear systems, a class that combines logical and switching devices together with more classical continuous variables. When the nonlinearities appear in the feedback loops, the problems become far more difficult; in that context, see, for instance, [11] for results about the computational power of systems of the type $x(t+1) = \sigma(Ax(t) + Bu(t))$.

1.1. Summary of paper. As mentioned earlier, the focus of this paper is the class of *sign-linear systems*, that is, those of the type (1) with $\sigma(x) = \text{sign}(x)$ (the sign is understood as being taken in each coordinate, so that the output value space could be taken simply as $\{-1, 0, 1\}^p$; careful definitions are given later). Also of interest are the associated *sign-linear input/output (i/o) maps* of the form

$$y(t) = \text{sign}(\mathcal{A}_1 u(t) + \cdots + \mathcal{A}_t u(1))$$

or the analogous continuous-time maps (convolution followed by sign).

For a system such as (1), we call any triple (A, B, C) such that the equations of the system can be expressed in terms of that triple, a *triple associated to the system* Σ . Note that in some cases there may be many triples associated to a single system:

Example 1.1. Let Σ be a sign-linear system with state-space \mathbb{R}^2 defined by the equations

$$\begin{aligned} x(t+1) &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u(t), \\ y(t) &= \text{sign}(x_1 + 2x_2). \end{aligned}$$

Then the following triples are both associated to Σ :

$$\begin{aligned} A &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad C = (1 \ 2), \\ A &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad C = (2 \ 4). \end{aligned}$$

The results that we describe parallel those known for standard linear systems, but with a few, perhaps unexpected, differences. We may summarize the main conclusions on sign-linear systems as follows:

(a) Although stronger than just observability of the pair (A, C) , observability of sign-linear systems can be characterized in an elegant manner. The characterization is different in the continuous- and discrete-time cases, in contrast to what happens for linear systems. Moreover, in another characteristic that is typical of *nonlinear* systems, the degree of controllability of the system does affect observability.

(b) Minimal-dimensional realizations of sign-linear i/o maps by sign-linear systems are unique up to a change of variables in the state space and a positive rescaling of outputs. This is basically as in the linear case, except for the obvious need to rescale. Moreover, finite-dimensional realizability can be characterized in the usual manner using Hankel matrices.

(c) If a realization of a given sign-linear i/o map is controllable and observable, in the usual sense of control theory, then it is minimal. Conversely, a minimal realization is necessarily final-state observable (that is, there is a control allowing for determination of the state at the end of the interval of application) but, in the discrete-time case, minimal realizations may not be observable. (In continuous-time, final-state and plain —initial-state— observability coincide.)

(d) Because of the possible lack of observability of minimal realizations, for some discrete-time sign-linear i/o maps it is the case that the abstract “canonical” realization, known to exist from automata-theoretic arguments, is not given by a sign-linear system. We discuss the canonical systems that result when minimal realizations are not observable, obtaining a description in terms of cascades of finite automata and linear systems.

The paper ends in § 6, where we show how some of the continuous-time observability results can be seen as consequences of the corresponding discrete-time results by sampling at appropriate frequencies.

Some of the results to be given can be stated in more generality, in terms of constrained-output systems as in (1), where σ is a fixed nonlinearity satisfying some (or all) of the following axioms:

1. $\text{sign}(\sigma(x)) = \text{sign}(x)$.
2. Finite precision sensor: $\sigma(x) = \text{constant}$ for $x \in (0, \varepsilon]$ and $x \in [-\varepsilon, 0)$, for some $\varepsilon > 0$.
3. Sensor saturation: $\sigma(x) = \text{constant}$ for $x > K > 0$ and $x < -K < 0$, for some $K > 0$.
4. $\sigma(x)$ is not constant on $(0, \infty)$ or $(-\infty, 0)$.

(Again, for a vector $z \in \mathbb{R}^p$, the notation $\sigma(z)$ denotes the vector $(\sigma(z_1), \dots, \sigma(z_p)) \in \mathbb{R}^p$.) The main systems of interest in this paper, sign-linear ones, are those for which $\sigma(x) = \text{sign}(x)$, which satisfies axioms 1,2,3. Some other examples of constrained-output systems are as follows:

- Output-saturated systems (satisfying 1,3,4) are those with $\sigma(x) = s(x)$, where

$$s(x) = \begin{cases} 1 & \text{if } x > 1, \\ x & \text{if } |x| \leq 1, \\ -1 & \text{if } x < -1. \end{cases}$$

The output space for output-saturated systems is $[-1, 1]^p$.

- Quantized systems (satisfying 1,2,4) are defined by $\sigma(x) = \lfloor x \rfloor$, with output space \mathbb{Z}^p .
- Saturated-quantized systems (satisfying 1–4) are systems for which

$$\sigma(x) = \begin{cases} K \text{sign}(x) & \text{if } |x| > K, \\ \lfloor x \rfloor & \text{if } |x| < K \end{cases}$$

for some fixed $K > 0$. Saturated-quantized systems have output space $\{n \in \mathbb{Z} : |n| \leq K\}^p$.

More details about such functions, and results specific for some of these classes, are described in [10].

2. Observability. Our notion of observability is the usual concept of multiple-experiment observability. Let us recall the main ideas. For formal definitions, please refer to [12, §5.1].

DEFINITION 2.1. A system Σ is *observable* if for any two initial states, there is some control that produces different outputs for each of the two initial states. This is not in general equivalent to single-experiment observability in which there exists one control function (or sequence of controls) that distinguishes any pair of states. The control we use to distinguish two states may depend on the two given states. This concept of observability really tells us only that we may *distinguish* between any two initial states, not that we may *determine* the initial state using one special control. For linear systems, multiple- and single-experiment observability are equivalent. If a linear system is observable, then the zero control will distinguish any pair of states.

DEFINITION 2.2. A system Σ is *final-state observable* if for any two initial states, there is some control and some time T so that either the output before time T is different for each of the two states, or the states at time T are the same. For continuous-time systems, final-state observability is equivalent to observability ([12, Prop. 5.1.9]).

We now state a few general necessary conditions for observability of constrained-output systems. Later, we will provide necessary and sufficient conditions for the class of sign-linear systems. The following result is obvious.

LEMMA 2.3. *If Σ is an observable constrained-output system, then (A, C) is an observable pair.*

Conversely, if σ is one-to-one, then observability of the pair (A, C) implies observability of Σ , but in general the implication does not hold. The following lemma gives an additional necessary condition when σ is not one-to-one.

LEMMA 2.4. *If Σ is an observable discrete-time constrained-output system with a single output channel ($p = 1$), and σ is not one-to-one, then $\det A \neq 0$.*

Proof. Suppose $\det A = 0$. Then there exists a nonzero $x \in \ker A$. The output sequence for the initial state x is $\{\sigma(Cx), \dots\}$ where the part not shown is independent of x . Since Σ is observable, (A, C) is an observable pair, so $Cx \neq 0$. Let $\sigma(\mu) = \sigma(\nu)$, $\mu \neq \nu$. Then, we may choose $\alpha_1 \neq \alpha_2$ so that $\alpha_1 Cx = \mu$ and $\alpha_2 Cx = \nu$. Then $\alpha_1 x \neq \alpha_2 x$ are indistinguishable, contradicting observability. \square

For $p > 1$, this lemma is not necessarily true. Consider the following counterexample. We will use the notation $x^+(t)$ to mean $x(t+1)$, and we drop the argument t from now on.

Example 2.5. Let Σ be the system with equations $x^+ = 0$, $y_1 = \sigma(x)$, $y_2 = \sigma(2x)$; and

$$\sigma(x) = \begin{cases} x & x \notin [1, 2] \\ 1 & x \in [1, 2] \end{cases}.$$

The nonlinearity σ is not one-to-one, but the map $x \mapsto (\sigma(x), \sigma(2x))$ is one-to-one, so the system is observable. However, A is not invertible.

If the measurement limiter σ is some form of saturation or σ has finite precision near 0, observability does imply that A is invertible, even in the multiple output case, since the following lemma will apply.

LEMMA 2.6. *If Σ is an observable discrete-time constrained-output system and σ either models sensor saturation*

$$\sigma(x) = \text{constant for } x > K > 0 \text{ and } x < -K < 0$$

for some $K > 0$, or has finite precision

$$\sigma(x) = \text{constant for } x \in (0, \varepsilon] \text{ and } x \in [-\varepsilon, 0)$$

for some $\varepsilon > 0$, then $\det A \neq 0$.

Proof. If $\det A = 0$, then there exists a nonzero $x \in \ker A$. In the saturated case, choose λ so that for all i satisfying $C_i x \neq 0$, then $|\lambda C_i x| > K$. Then λx and $2\lambda x$ are indistinguishable. In the finite precision case, choose λ so that $|\lambda C_i x| \leq \varepsilon$ for all $i = 1, \dots, p$. Then λx and $\frac{1}{2}\lambda x$ are indistinguishable. \square

Before stating the next lemma we introduce the following definition.

DEFINITION 2.7. Let Σ be a constrained-output system and let (A, B, C) be any triple associated to Σ . Then the sequence of $p \times m$ matrices

$$\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \dots\},$$

where

$$\mathcal{A}_i = CA^{i-1}B, \quad i = 1, 2, 3, \dots$$

is called a *Markov parameter sequence* associated to Σ . Since in general C is not uniquely defined from the system equations, there may be more than one Markov sequence associated to a given system; this issue is discussed later.

LEMMA 2.8. Assume that Σ is a single-output observable discrete-time constrained-output system defined by the triple (A, B, C) , and σ has finite precision

$$\sigma(x) = \text{constant for } x \in (0, \varepsilon] \text{ and } x \in [-\varepsilon, 0).$$

If A has an eigenvalue λ satisfying $|\lambda| \leq 1$, then $\mathcal{A} \not\equiv 0$, for any Markov sequence associated to Σ .

Proof. Let \mathcal{A} be any Markov sequence associated to Σ . Let v be a nonzero eigenvector for A corresponding to λ and let $\gamma = \|v\|$ (where $\|\cdot\|$ denotes Euclidean norm). Then $A^k v = \lambda^k v$ for all k , so $\|A^k v\| \leq \gamma$ for all k . Write $v = v_1 + iv_2$, where v_1, v_2 are real vectors. Then $\|A^k v_1\| \leq \|A^k v\| \leq \gamma$ for all k . Note that $\|C\| \neq 0$ since (A, C) is an observable pair. Then

$$x := \frac{v_1 \varepsilon}{\gamma \|C\|}$$

satisfies

$$|CA^k x| \leq \|C\| \|A^k x\| \leq \frac{\|C\| \gamma \varepsilon}{\gamma \|C\|} = \varepsilon$$

for all k . If $\mathcal{A} \equiv 0$, then $x, \frac{1}{2}x$ are indistinguishable, contradicting observability. \square

3. Sign-linear systems. Now we concentrate on the observability of sign-linear systems. Sign-linear input/output maps and their realizations will be discussed in §§ 4 and 5.

DEFINITION 3.1. A *sign-linear system* Σ is a system with state, input, and output-value spaces \mathbb{R}^n , \mathbb{R}^m , and $\{-1, 0, 1\}^p$, respectively, for which there exist matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$, so that the equations of Σ take the form

$$\begin{aligned} x^+ (\text{or } \dot{x}) &= Ax + Bu, \\ y &= \text{sign}(Cx) \end{aligned}$$

in discrete- (or continuous-) time. If (A, B, C) is a triple like this, we denote $\Sigma = (A, B, C)_s$. Whether we are dealing with discrete- or continuous-time will be clear

from the context. The integer n is the *dimension* of the system. It is convenient to include the degenerate case $n = 0$, corresponding to the system with zero-dimensional state space.

Note that $(A, B, C)_s = (\hat{A}, \hat{B}, \hat{C})_s$ if and only if $A = \hat{A}$, $B = \hat{B}$, and $C = \Lambda \hat{C}$, where Λ is a *scaling matrix* in the following sense.

DEFINITION 3.2. A $p \times p$ *scaling matrix* is a matrix of the type

$$\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p) = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix},$$

where $\lambda_i > 0$, $i = 1, \dots, p$. Any triple for which $\Sigma = (A, B, C)_s$ will be said to be *associated* to Σ . Observe that the properties of (A, B) being a controllable pair, (A, C) being an observable pair, and (A, B, C) being canonical (controllable and observable), in the usual linear systems sense, are independent of which of the associated triples is considered.

The following trivial observation will be used often.

Remark 3.3. If \mathcal{H} is a real pre-Hilbert space (that is, a space with a nondegenerate inner product), and if $c \in \mathcal{H}$, c nonzero, $a, b \in \mathbb{R}$, $a \neq b$, then there is a $u \in \mathcal{H}$ so that

$$\text{sign}(a + \langle c, u \rangle) \neq \text{sign}(b + \langle c, u \rangle).$$

Indeed, without loss of generality, we may assume that $a > b$. Let

$$\alpha := -\frac{a+b}{2\|c\|^2}.$$

Then $u := \alpha c$ satisfies $a + \langle c, u \rangle > 0$ and $b + \langle c, u \rangle < 0$.

Since controllability of a system does not depend on outputs, a sign-linear system is controllable if and only if (A, B) is a controllable pair in the usual sense ([12]). Observability requires a bit more than in the linear case, as illustrated by the system

$$x(t+1) = u(t), \quad y(t) = \text{sign}(x(t)).$$

Observability of the pair (A, C) is not sufficient to guarantee observability of the corresponding sign-linear system.

We will say that a triple (A, B, C) has *property \mathcal{P}* if not only is (A, C) observable, but we can choose a subset of outputs which allow observability of the pair (A, C) and for each of which the corresponding row of the Markov sequence \mathcal{A} is nonzero. In the case $p = 1$, this just means that (A, C) is an observable pair and $\mathcal{A} \neq 0$, or equivalently, that (A, C) is an observable pair and $B \neq 0$. For $p > 1$, \mathcal{A} has p rows and we only require that enough of those rows are nonzero. More precisely, let

$$I(\mathcal{A}) = \{i_1, \dots, i_k\}$$

be the indices of the nonzero rows of \mathcal{A} ; then property \mathcal{P} is the condition that

$$(3) \quad \bigcap_{\substack{j \in I(\mathcal{A}) \\ q=0, \dots, n-1}} \ker(C_j A^q) = \{0\},$$

where C_j denotes the j th row of C . Note that if C and \hat{C} differ only by multiplication by a scaling matrix, property \mathcal{P} holds for (A, B, C) if and only if it holds for (A, B, \hat{C}) .

Thus there is no ambiguity in the following statements. For discrete- and continuous-time the following theorems state necessary and sufficient conditions for observability.

THEOREM 3.4. *Let $\Sigma = (A, B, C)_s$ be a sign-linear discrete-time system of dimension $n > 0$. Then, Σ is observable if and only if the following conditions hold:*

1. $\det A \neq 0$,
2. (A, B, C) has property \mathcal{P} .

Proof. Necessity. Suppose Σ is observable. We know $\det A \neq 0$ from Lemma 2.6. Now assume that property \mathcal{P} would not hold, and pick $x \neq 0$ in the intersection in (3). The output sequence for any given control sequence $\{u_1, u_2, \dots\}$ is $\{y(0), y(1), \dots\}$ where

$$y(k) = \text{sign} \left(CA^k x + \sum_{l=1}^k \mathcal{A}_l u_{k-l+1} \right).$$

For the chosen x in that intersection, each row of each term in the output sequence has the form

$$y(k)_j = \begin{cases} \text{sign}(C_j A^k x + 0) & \text{if } j \notin I(\mathcal{A}), \\ \text{sign}(0 + *) & \text{if } j \in I(\mathcal{A}), \end{cases}$$

where $*$ denotes a (possibly nonzero) function of the inputs and the Markov parameters. Then, x and λx for any $\lambda > 0$, $\lambda \neq 1$, cannot be distinguished, so observability is contradicted.

Sufficiency. Now suppose $\det A \neq 0$ and (A, B, C) has property \mathcal{P} . We must show that Σ is observable. Pick an integer $l > 0$ so that the i th row of

$$(\mathcal{A}_1 \mathcal{A}_2 \cdots \mathcal{A}_l)$$

is nonzero for every $i \in I(\mathcal{A})$. Note that since A is invertible,

$$(4) \quad \bigcap_{\substack{j \in I(\mathcal{A}) \\ q=0, \dots, n-1}} \ker(C_j A^{q+l}) = \{0\},$$

which follows from (3).

Now look at the following n terms in the output sequence for initial state x :

$$\begin{aligned} & \text{sign}(CA^l x + \mathcal{A}_l u_1 + \cdots + \mathcal{A}_1 u_l), \\ & \text{sign}(CA^{l+1} x + \mathcal{A}_{l+1} u_1 + \cdots + \mathcal{A}_1 u_{l+1}), \dots, \\ & \text{sign}(CA^{l+n-1} x + \mathcal{A}_{l+n-1} u_1 + \cdots + \mathcal{A}_1 u_{l+n-1}). \end{aligned}$$

Given $x \neq z$ we must show that x, z are distinguishable. If we can choose a sequence $u_1, u_2, \dots, u_{l+n-1}$ so that some row of some term above is different for the initial states x and z , then x, z are distinguishable. As $x - z \neq 0$, we may pick some $j \in I(\mathcal{A})$ and some $q = 0, \dots, n-1$ so that

$$C_j A^{q+l} x \neq C_j A^{q+l} z.$$

Since $j \in I(\mathcal{A})$, the j th row of $(\mathcal{A}_1 \cdots \mathcal{A}_l)$ is nonzero by our choice of l . Denote $k := q + l$ so that the j th row of $(\mathcal{A}_1 \cdots \mathcal{A}_k)$ is also nonzero. Let \mathcal{A}_i^j be the j th row of

\mathcal{A}_i . Then we may apply Remark 3.3 (with $\mathcal{H} = \mathbb{R}^k$ and the standard inner product) and obtain u_1, u_2, \dots, u_k so that

$$\begin{aligned} & \text{sign}(C_j A^k x + \mathcal{A}_k^j u_1 + \dots + \mathcal{A}_1^j u_k) \\ & \neq \text{sign}(C_j A^k z + \mathcal{A}_k^j u_1 + \dots + \mathcal{A}_1^j u_k). \end{aligned}$$

Thus, x and z are distinguishable. This completes the proof. \square

We will say that a triple (A, B, C) is *discrete-time sign-linear observable* if the triple satisfies the observability conditions in Theorem 3.4.

For continuous-time sign-linear systems, the conditions for observability are slightly weaker, as invertibility of the matrix A is not needed.

THEOREM 3.5. *Let $\Sigma = (A, B, C)_s$ be a sign-linear continuous-time system of dimension $n > 0$. Then Σ is observable if and only if (A, B, C) has property \mathcal{P} .*

Proof. The proof is exactly the same as in the discrete-time case. Indeed, if (3) is not satisfied and $x \neq 0$ is in the intersection of the kernels, consider the output

$$y(t) = \text{sign} \left(C e^{At} x + \int_0^t \sum_{k=1}^{\infty} \frac{\mathcal{A}_k(t-s)^{k-1}}{(k-1)!} u(s) ds \right).$$

Each row has the form

$$y(t)_j = \begin{cases} \text{sign}(C_j e^{At} x + 0) & \text{if } j \notin I(\mathcal{A}), \\ \text{sign}(0 + *) & \text{if } j \in I(\mathcal{A}), \end{cases}$$

where $*$ denotes a (possibly nonzero) function of the inputs and the Markov parameters. Then x and λx for any $\lambda > 0$, $\lambda \neq 1$, are indistinguishable, contradicting observability.

Now suppose (A, B, C) satisfies property \mathcal{P} . We must show that Σ is observable. Look at the output function for initial state x :

$$y(t) = \text{sign} \left(C e^{At} x + \int_0^t K(t-s) u(s) ds \right),$$

where

$$K(t-s) := \sum_{k=1}^{\infty} \frac{\mathcal{A}_k(t-s)^{k-1}}{(k-1)!}.$$

Given $x \neq z$ we must show that x, z are distinguishable. If we can choose a t and a control function $u(\cdot)$ of length t so that some row of $y(t)$ is different for the initial states x and z , then x, z are distinguishable. As $x - z \neq 0$, we may pick some $j \in I(\mathcal{A})$ and some $t \geq 0$ so that

$$C_j e^{At} x \neq C_j e^{At} z,$$

by property \mathcal{P} . Since $C_j e^{At} x$ is an analytic function of t , this is true in a neighborhood of $t = 0$ so we may, in fact, fix a $t > 0$ so that the inequality holds.

Next note that since $j \in I(\mathcal{A})$, $\mathcal{A}^j \neq 0$ so also $K_j(\cdot) \neq 0$. Now apply Remark 3.3 with $a = C_j e^{At} x$, $b = C_j e^{At} z$, $\mathcal{H} = \mathcal{L}^\infty[0, t]$ with the \mathcal{L}^2 inner product

$$\langle v(\cdot), u(\cdot) \rangle := \int_0^t v(s) u(s) ds,$$

and $c = K_j(t-s) \in \mathcal{H}$. Thus we may choose a measurable essentially bounded $u(\cdot)$ so that

$$\text{sign} \left(Ce^{At}x + \int_0^t K(t-s)u(s)ds \right) \neq \text{sign} \left(Ce^{At}z + \int_0^t K(t-s)u(s)ds \right).$$

This $u(\cdot)$ distinguishes x, z and the proof is complete. \square

4. Sign-linear realizations. We now focus on questions of realizability for the class of sign-linear systems. As we mentioned earlier, a sign-linear system does not have a *unique* associated Markov sequence. However, for sign-linear systems, we have the following obvious fact.

Remark 4.1. A Markov parameter sequence associated to $\Sigma = (A, B, C)_s$ is any sequence of $p \times m$ matrices $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \dots\}$ so that

$$(5) \quad \mathcal{A}_i = \Lambda C A^{i-1} B, \quad i = 1, 2, 3, \dots$$

for some scaling matrix Λ .

For the degenerate system, its (only) associated sequence is $\mathcal{A} \equiv 0$. If \mathcal{A} is associated to Σ , we also say that Σ *realizes* \mathcal{A} . If (A, B, C) is a triple of matrices and $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \dots\}$ is a Markov sequence so that $\mathcal{A}_i = C A^{i-1} B$ holds, we will say that (A, B, C) is a *linear representation* of \mathcal{A} . The standard terminology is “realization” (as a linear system), but this can lead to confusion here, since we are interested in sign-linear realizations. Note that the above definitions imply that for any given triple (A, B, C) , and any sequence of $p \times m$ matrices \mathcal{A} , the sign-linear system $\Sigma = (A, B, C)_s$ realizes \mathcal{A} if and only if $(A, B, \Lambda C)$ is a linear representation of \mathcal{A} for some scaling matrix Λ . In other words, there must exist a triple associated to Σ that represents \mathcal{A} .

The matrix

$$H_{s,t} = \begin{pmatrix} \mathcal{A}_1 & \cdots & \mathcal{A}_t \\ \vdots & \ddots & \vdots \\ \mathcal{A}_s & \cdots & \mathcal{A}_{s+t-1} \end{pmatrix}$$

is called the $s \times t$ *Hankel matrix* for the Markov sequence \mathcal{A} . The (*Hankel*) *rank* of a sequence \mathcal{A} is defined to be

$$\sup_{s,t} \text{rank } H_{s,t}.$$

An i/o map (for a precise definition, see [12, Rem. 2.2.2], is a function of controls u defined on some time interval $[\sigma, \tau]$, which gives the entire output function for the time interval $[\sigma, \tau]$.

DEFINITION 4.2. A $(p \times m)$ *discrete-time sign-linear i/o map* α is a discrete-time i/o map for which there exists some sequence of $(p \times m)$ matrices $\mathcal{A}_1, \mathcal{A}_2, \dots$, so that

$$(6) \quad \alpha(u)(j) = \text{sign}(\mathcal{A}_j u_1 + \cdots + \mathcal{A}_1 u_j)$$

for each input sequence $\{u_1, u_2, u_3, \dots\}$. A *continuous-time sign-linear i/o map* is a continuous-time i/o map α for which there exists an analytic kernel $K(t)$ with expansion

$$(7) \quad K(t) = \sum_{i=1}^{\infty} \mathcal{A}_i \frac{t^{i-1}}{(i-1)!}$$

so that

$$(8) \quad \alpha(u)(t) = \text{sign} \left(\int_0^t K(t-s)u(s)ds \right)$$

for every measurable essentially bounded control function $u(\cdot)$. In either case, any sequence of matrices $\mathcal{A}_1, \mathcal{A}_2, \dots$ as above is called a *Markov sequence* of the map α . We will study realizations of these i/o maps by sign-linear systems. It will be helpful to have a simple example in mind as we go through the definitions and results.

Example 4.3. Let $\mathcal{A} = \{1, -1, 1, -1, \dots\}$. Then α is a 1×1 discrete-time i/o map, where

$$\alpha(u)(j) = \text{sign} \left((-1)^{j-1}u_1 + \dots + u_{j-2} - u_{j-1} + u_j \right).$$

For the control $u = \{1, 0, 0, \dots\}$, $\alpha(u)(j) = (-1)^{j-1}$ for all j . For the control $u = \{1, 2, 3, 0, 0, 0, \dots\}$, the values of the i/o function are

$$\begin{aligned} \alpha(u)(1) &= \text{sign}(1) = 1, \\ \alpha(u)(2) &= \text{sign}(-1 + 2) = 1, \\ \alpha(u)(3) &= \text{sign}(1 - 2 + 3) = 1, \\ \alpha(u)(4) &= \text{sign}(-1 + 2 - 3) = -1, \\ \alpha(u)(j) &= (-1)^{j-1} \text{ for } j > 4. \end{aligned}$$

DEFINITION 4.4. Two triples (A, B, C) and $(\hat{A}, \hat{B}, \hat{C})$ are *sign-similar* if there is a $T \in Gl(n)$ and a scaling matrix Λ such that

$$\begin{aligned} T^{-1}AT &= \hat{A}, \\ T^{-1}B &= \hat{B}, \\ CT &= \Lambda\hat{C}. \end{aligned}$$

If $\Sigma = (A, B, C)_s$ and $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C})_s$ are sign-linear systems, they are called *sign-similar* if the corresponding triples are. Note that sign-similarity is an equivalence relation, and that the ambiguity in defining a triple associated to Σ causes no difficulties in the above definition.

DEFINITION 4.5. Two Markov sequences

$$\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \dots\}, \quad \text{and} \quad \hat{\mathcal{A}} = \{\hat{\mathcal{A}}_1, \hat{\mathcal{A}}_2, \dots\}$$

are *sign-equivalent* if there exists a scaling matrix Λ so that

$$\mathcal{A}_j = \Lambda\hat{\mathcal{A}}_j, \quad j = 1, 2, 3, \dots$$

Note that if $K(\cdot)$ and $\hat{K}(\cdot)$ are as in (7) for the sequences \mathcal{A} and $\hat{\mathcal{A}}$, sign-equivalence of \mathcal{A} and $\hat{\mathcal{A}}$ is the same as asking that $K(t) = \Lambda\hat{K}(t)$ for all t , where Λ is a scaling matrix. The Markov sequence \mathcal{A} in Example 4.3 is sign-equivalent to $\hat{\mathcal{A}} = \{a, -a, a, -a, \dots\}$ for any $a > 0$.

4.1. Basic facts about realizations. The next lemma says that the Markov sequence \mathcal{A} is uniquely determined by a sign-linear i/o map α up to multiplication by a scaling matrix. That is, a sign-linear i/o map is defined by many Markov sequences, but these sequences are related by scaling.

Observe that the impulse response of a sign-linear i/o map (e.g., for discrete-time systems, the response to the input $u = \{1, 0, 0, 0, \dots\}$) is not enough to uniquely characterize the i/o map. For a discrete-time sign-linear i/o map α , the impulse response is just the sequence of signs of the Markov parameters: $\{\text{sign}(\mathcal{A}_1), \text{sign}(\mathcal{A}_2), \dots\}$. Such a sign sequence represents infinitely many different families of sign-equivalent Markov sequences, as illustrated by the following example.

Example 4.6. Let α_1 be the discrete-time sign-linear i/o map defined by $\mathcal{A} = \{1, 3, 1, 3, \dots\}$ and $\hat{\alpha}$ the map defined by $\hat{\mathcal{A}} = \{3, 1, 3, 1, \dots\}$. Then the impulse response for both i/o maps is $\{+1, +1, +1, \dots\}$; however the two maps are not the same as shown by considering the output corresponding to the input $u = \{1, -1, 1, -1, \dots\}$:

$$\begin{aligned}\alpha(u)(1) &= +1; & \hat{\alpha}(u)(1) &= +1, \\ \alpha(u)(2) &= +1; & \hat{\alpha}(u)(2) &= -1, \\ \alpha(u)(3) &= -1; & \hat{\alpha}(u)(3) &= +1, \\ \alpha(u)(4) &= +1; & \hat{\alpha}(u)(4) &= -1.\end{aligned}$$

LEMMA 4.7. \mathcal{A} and $\hat{\mathcal{A}}$ define the same i/o map if and only if they are sign-equivalent.

Proof. If \mathcal{A} and $\hat{\mathcal{A}}$ are sign-equivalent, then $\mathcal{A}_i = \Lambda \hat{\mathcal{A}}_i$ for all i , where Λ is a scaling matrix. It is then clear from (6)–(8) that the corresponding i/o maps coincide. To prove the converse, we can assume, without loss of generality, looking at each component of the output and each row of \mathcal{A} , that $p = 1$. We first prove the following easy observation.

Remark 4.8. If \mathcal{V} is a real pre-Hilbert space and if $v, w \in \mathcal{V}$ are nonzero and such that

$$\text{sign} \langle v, u \rangle = \text{sign} \langle w, u \rangle$$

for all $u \in \mathcal{V}$, then there exists $\lambda > 0$ so that $v = \lambda w$.

Proof. Suppose first that v and w are linearly independent, and consider the plane they span. Let $u \neq 0$ be in this plane and perpendicular to $v + w$. As $\langle v + w, u \rangle = 0$, necessarily $\langle v, u \rangle \neq 0$ and $\langle w, u \rangle \neq 0$, since if either of these is zero, then the other one is too, and that would contradict linear independence. Then

$$\langle v, u \rangle + \langle w, u \rangle = \langle v + w, u \rangle = 0.$$

So $\langle v, u \rangle = -\langle w, u \rangle \neq 0$, contradicting the assumption. Thus, either $v = \lambda w$ with $\lambda > 0$, or $v = -\mu w$ with $\mu > 0$. If $v = -\mu w$, then

$$\langle v, u \rangle = \langle -\mu w, u \rangle = -\mu \langle w, u \rangle \neq 0$$

and so $\text{sign} \langle v, u \rangle \neq \text{sign} \langle w, u \rangle$, again a contradiction. The only remaining possibility is that there exists a $\lambda > 0$ so that $v = \lambda w$. \square

Now we can continue the proof of Lemma 4.7. For discrete-time, we must show that if

$$\text{sign} \left(\sum_{i=1}^l \mathcal{A}_i u_{l-i+1} \right) = \text{sign} \left(\sum_{i=1}^l \hat{\mathcal{A}}_i u_{l-i+1} \right)$$

for all $l \geq 1$ and for all u_1, u_2, \dots, u_l , then $\mathcal{A} = \lambda \hat{\mathcal{A}}$ for some $\lambda > 0$. First choose an l so that $(\mathcal{A}_1, \dots, \mathcal{A}_l) \neq 0$. Note that $\mathbb{R}^{lm} = (\mathbb{R}^m)^l$ forms a pre-Hilbert space with the standard inner product

$$\left\langle \begin{pmatrix} v_1 \\ \vdots \\ v_l \end{pmatrix}, \begin{pmatrix} u_1 \\ \vdots \\ u_l \end{pmatrix} \right\rangle := \sum_{i=1}^l v'_i u_i.$$

Applying Remark 4.8 with $v = (\mathcal{A}_1, \dots, \mathcal{A}_l)'$, $w = (\hat{\mathcal{A}}_1, \dots, \hat{\mathcal{A}}_l)'$ and $u = (u'_1, \dots, u'_l)'$, we see that there exists a $\lambda > 0$ so that

$$(\mathcal{A}_1, \dots, \mathcal{A}_l) = \lambda(\hat{\mathcal{A}}_1, \dots, \hat{\mathcal{A}}_l).$$

Now pick any $q > l$. Applying the same argument to $(\mathbb{R}^m)^q$, we obtain a $\lambda_q > 0$ so that

$$(\mathcal{A}_1, \dots, \mathcal{A}_q) = \lambda_q(\hat{\mathcal{A}}_1, \dots, \hat{\mathcal{A}}_q).$$

Since $(\mathcal{A}_1, \dots, \mathcal{A}_l)$ is a subvector of $(\mathcal{A}_1, \dots, \mathcal{A}_q)$, and similarly $(\hat{\mathcal{A}}_1, \dots, \hat{\mathcal{A}}_l)$ a subvector of $(\hat{\mathcal{A}}_1, \dots, \hat{\mathcal{A}}_q)$, this implies that $\lambda = \lambda_q$. Thus $\mathcal{A}_q = \lambda \hat{\mathcal{A}}_q$ for all $q \geq 1$.

For continuous-time, we need to show that if

$$\text{sign} \left(\int_0^t K(t-s)u(s)ds \right) = \text{sign} \left(\int_0^t \hat{K}(t-s)u(s)ds \right)$$

for all $t \in [0, \infty)$ and for all $u(\cdot)$, measurable and essentially bounded on $[0, t]$, then $K(t) = \lambda \hat{K}(t)$ for some $\lambda > 0$ and for all $t \geq 0$. Note that $\mathcal{L}^\infty[0, t]$ forms a pre-Hilbert space with the \mathcal{L}^2 inner product

$$\langle v(\cdot), u(\cdot) \rangle := \int_0^t v(s)u(s)ds.$$

Applying Remark 4.8 with $v(s) = K(t-s)$ and $w(s) = \hat{K}(t-s)$, we see that there exists a $\lambda_t > 0$ so that $K(\cdot)|_{[0,t]} = \lambda_t \hat{K}(\cdot)|_{[0,t]}$. Using an argument similar to the one used in the discrete-time case, we can conclude that there exists a $\lambda > 0$ so that $K(t) = \lambda \hat{K}(t)$ for all $t \geq 0$. \square

COROLLARY 4.9. *Let α be a sign-linear i/o map, with Markov sequence \mathcal{A} , and let $\Sigma = (A, B, C)_s$ be any sign-linear realization of α , with Markov sequence $\hat{\mathcal{A}}$. Then \mathcal{A} and $\hat{\mathcal{A}}$ are sign-equivalent.*

Proof. Just note that \mathcal{A} and $\hat{\mathcal{A}}$ define the same i/o map, namely α . Thus, the previous lemma applies. \square

4.2. Minimality.

DEFINITION 4.10. A sign-linear system of dimension n is *minimal* if any other sign-linear system realizing the same i/o map has dimension $n_1 \geq n$. Recall that a triple (A, B, C) is *canonical* if and only if it is a minimal-dimensional linear representation of its Markov sequence ([12, Thm. 20]). The next lemma states that minimality of a sign-linear system is equivalent to minimality of the associated linear system.

LEMMA 4.11. *The sign-linear system $(A, B, C)_s$ is a minimal realization of α if and only if the triple (A, B, C) is canonical.*

Proof. Suppose the sign-linear system $\Sigma = (A, B, C)_s$ is a minimal realization of the i/o map α and \mathcal{A} is a Markov sequence associated to α . If (A, B, C) is not canonical, then there exists another triple $(\hat{A}, \hat{B}, \hat{C})$ of smaller dimension that is a linear representation of the same Markov sequence $\hat{\mathcal{A}}$ as (A, B, C) . Then $\Sigma = (A, B, C)_s$ also realizes $\hat{\mathcal{A}}$ so $\hat{\mathcal{A}}$ is sign-equivalent to \mathcal{A} by Corollary 4.9. But then since $(\hat{A}, \hat{B}, \hat{C})_s$ realizes $\hat{\mathcal{A}}$, it also realizes \mathcal{A} (a Markov sequence associated to a sign-linear system is only determined up to sign-equivalence). Thus, $(\hat{A}, \hat{B}, \hat{C})_s$ is a sign-linear system realizing α and of smaller dimension than $(A, B, C)_s$, contradicting minimality.

Conversely, suppose the triple (A, B, C) is canonical of dimension n , which implies, in particular, that it is minimal. If $(A, B, C)_s$ is not also minimal, then there exists a sign-linear system $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C})_s$ of dimension $n_1 < n$ that realizes the same i/o map α as $(A, B, C)_s$. Let \mathcal{A} be the Markov sequence represented by (A, B, C) and $\hat{\mathcal{A}}$ the sequence represented by $(\hat{A}, \hat{B}, \hat{C})$. Then $(A, B, C)_s$ realizes \mathcal{A} and $(\hat{A}, \hat{B}, \hat{C})_s$ realizes $\hat{\mathcal{A}}$. Since the two sign-linear systems realize the same i/o map α , \mathcal{A} and $\hat{\mathcal{A}}$ are sign-equivalent (Corollary 4.9). Thus, there exists a scaling matrix Λ so that $\mathcal{A}_j = \Lambda \hat{\mathcal{A}}_j$ for all $j \geq 1$. So

$$CA^{i-1}B = \Lambda \hat{C} \hat{A}^{i-1} \hat{B}, \quad i = 1, 2, 3, \dots$$

But then $(\hat{A}, \hat{B}, \Lambda \hat{C})$ is a linear representation of \mathcal{A} of dimension $n_1 < n$, contradicting the minimality of (A, B, C) . \square

Remark 4.12. If \mathcal{A} and $\hat{\mathcal{A}}$ are two Markov sequences associated to the same i/o map α , then $\text{rank } \mathcal{A} = \text{rank } \hat{\mathcal{A}}$. Thus, we can define the *Hankel rank* of α as the rank of any of the associated Markov sequences. Indeed, by Lemma 4.7, we know that \mathcal{A} and $\hat{\mathcal{A}}$ are sign-equivalent. Thus there is a scaling matrix Λ with $\mathcal{A}_j = \Lambda \hat{\mathcal{A}}_j$, $j = 1, 2, 3, \dots$. We then have, for any $s, t \geq 1$,

$$H_{s,t} = \Lambda_s \hat{H}_{s,t},$$

where $\hat{H}_{s,t}$ is the $s \times t$ Hankel matrix for $\hat{\mathcal{A}}$ and $\Lambda_s = \text{diag}(\Lambda, \dots, \Lambda)$. Since this is true for every s, t ,

$$\begin{aligned} \text{rank}(\mathcal{A}) &= \sup_{s,t} \{\text{rank}(H_{s,t})\} \\ &= \sup_{s,t} \{\text{rank}(\hat{H}_{s,t})\} = \text{rank}(\hat{\mathcal{A}}), \end{aligned}$$

as claimed.

THEOREM 4.13. *Let α be a sign-linear i/o map. Then α is realizable by a sign-linear system if and only if α has finite Hankel rank.*

Proof. If α has finite rank then any Markov sequence for α , \mathcal{A} , has finite rank. It then follows that there exists a linear representation for \mathcal{A} , (A, B, C) . Then the corresponding sign-linear system $(A, B, C)_s$ realizes α .

Conversely, given a sign-linear i/o map α that is realizable by a sign-linear system $(A, B, C)_s$, we would like to show that α has finite rank. One Markov sequence for $(A, B, C)_s$ is the impulse response of the linear system (A, B, C) . This impulse response \mathcal{A} is a Markov sequence for α . From linear realization theory, we know that \mathcal{A} has finite rank. Thus, by the remark above, α has finite rank. \square

LEMMA 4.14. *If (A, B, C) is a canonical representation of a Markov sequence \mathcal{A} , then (A, B, C) satisfies property \mathcal{P} .*

Proof. Suppose (A, B, C) does not satisfy property \mathcal{P} . Then by observability there is some $i \notin I(\mathcal{A})$ (i.e., so that the i th row \mathcal{A}^i of \mathcal{A} is zero) so that

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ l=0, \dots, n-1}} \ker(C_j A^l) \neq \{0\},$$

but

$$(9) \quad \bigcap_{\substack{j \in I(\mathcal{A}) \cup \{i\} \\ l=0, \dots, n-1}} \ker(C_j A^l) \subsetneq \bigcap_{\substack{j \in I(\mathcal{A}) \\ l=0, \dots, n-1}} \ker(C_j A^l).$$

Since $\mathcal{A}^i \equiv 0$ then $C_i(A^j B) = 0$ for all j , so in particular,

$$C_i \begin{pmatrix} B & AB & A^2 B & \dots & A^{n-1} B \end{pmatrix} = 0.$$

The pair (A, B) is controllable so

$$\begin{pmatrix} B & AB & A^2 B & \dots & A^{n-1} B \end{pmatrix}$$

has full row rank. Thus $C_i = 0$, contradicting (9). So property \mathcal{P} indeed holds. \square

LEMMA 4.15. *If (A, B, C) is a triple satisfying property \mathcal{P} , then the sign-linear system $\Sigma = (A, B, C)_s$ is final-state observable.*

Proof. First suppose $\Sigma = (A, B, C)_s$ is a discrete-time sign-linear system. Perform a change of variables in the state space. Let

$$z = T^{-1}x = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where $T \in Gl(n)$ is chosen so that

$$T^{-1}AT = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix},$$

with A_1 of size $n_1 \times n_1$ nilpotent and A_2 of size $n_2 \times n_2$ nonsingular. (This can be done, for instance, by first putting A in real canonical form and then reordering the blocks so that the blocks corresponding to 0 eigenvalues come first.) Then

$$y = \text{sign}(CTz)$$

can be written as

$$y = \text{sign} \left[\begin{pmatrix} C_1 & C_2 \end{pmatrix} z \right],$$

and we can also write

$$T^{-1}B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

Since (A, C) is an observable pair, the n columns of

$$\begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = \begin{pmatrix} C_1 & C_2 \\ C_1 A_1 & C_2 A_2 \\ \vdots & \vdots \\ C_1 A_1^{n-1} & C_2 A_2^{n-1} \end{pmatrix} T^{-1}$$

are linearly independent. As T^{-1} is an invertible matrix, both (A_1, C_1) and (A_2, C_2) must be observable pairs. Property \mathcal{P} implies that the subset of outputs indexed by $I(\mathcal{A})$ allows observability of the pair (A, C) . Then the outputs indexed by $I(\mathcal{A})$ also allow observability of the pair (A_2, C_2) . We know that $\mathcal{A}^i \neq 0$ for $i \in I(\mathcal{A})$. Since A_1 is nilpotent, after $n + 1$ steps the output sequence looks like

$$\begin{aligned} & \text{sign}(C_2 A_2^{n+1} z_2 + \mathcal{A}_1 u_{n+1} + \cdots + \mathcal{A}_{n+1} u_1), \\ & \text{sign}(C_2 A_2^{n+2} z_2 + \mathcal{A}_1 u_{n+2} + \cdots + \mathcal{A}_{n+2} u_1), \dots \end{aligned}$$

Now we have (A_2, C_2) is an observable pair, $\det A_2 \neq 0$, and

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ q=0, \dots, n_2-1}} \ker((C_2)_j A_2^q) = \{0\},$$

where $\mathcal{A}^i \neq 0$ for $i \in I(\mathcal{A})$. Now using Remark 3.3, we may always choose appropriate controls to distinguish any distinct z_2 and \tilde{z}_2 . Also, z_1 goes to zero (in less than n time steps). So the system is final-state observable.

For a continuous-time sign-linear system, $\Sigma = (A, B, C)_s$, property \mathcal{P} alone implies that the system is observable, by Lemma 3.5. Hence, Σ is also final-state observable. (Observability and final-state observability are equivalent in continuous-time.) \square

THEOREM 4.16.

1. *If a sign-linear realization is controllable and observable then it is minimal.*
2. *If it is minimal then it is controllable and final-state observable.*
3. *Any two minimal sign-linear realizations are sign-similar.*

Proof. 1. If the sign-linear system $(A, B, C)_s$ is controllable and observable then in particular the triple (A, B, C) is canonical so the sign-linear system is minimal by Lemma 4.11.

2. If the system $\Sigma = (A, B, C)_s$ is minimal, then the triple (A, B, C) is canonical. If $\mathcal{A} \equiv 0$, then the minimal realization has dimension 0 and is trivially final-state observable. So now assume that we are dealing with dimension $n > 0$. We know that the triple (A, B, C) is canonical, so it satisfies property \mathcal{P} (Lemma 4.14). Next, applying Lemma 4.15, we conclude that $\Sigma = (A, B, C)_s$ is final-state observable.

3. Given two minimal realizations $(A, B, C)_s$ and $(\hat{A}, \hat{B}, \hat{C})_s$, of a sign-linear map α , with Markov sequence \mathcal{A} , we must show that they are sign-similar. The corresponding triples (A, B, C) and $(\hat{A}, \hat{B}, \hat{C})$ represent Markov sequences \mathcal{A}^1 and \mathcal{A}^2 , respectively, which are both sign-equivalent to \mathcal{A} (Corollary 4.9). That is, we have scaling matrices Λ_1, Λ_2 satisfying

$$\mathcal{A} = \Lambda_1 \mathcal{A}^1, \quad \mathcal{A} = \Lambda_2 \mathcal{A}^2.$$

Since the sign-linear realizations are minimal, the linear representations are canonical (Lemma 4.11). Since Λ_1 and Λ_2 have full rank, this implies that $(A, \Lambda_1 C)$ and $(\hat{A}, \Lambda_2 \hat{C})$ are also observable pairs. Thus, $(A, B, \Lambda_1 C)$ and $(\hat{A}, \hat{B}, \Lambda_2 \hat{C})$ are both canonical linear representations of the same Markov sequence \mathcal{A} . By [12], Thm. 20, they must be similar, i.e., there exists some $T \in Gl(n)$ so that $T^{-1}AT = \hat{A}$, $T^{-1}B = \hat{B}$, and $\Lambda_1 CT = \Lambda_2 \hat{C}$. Thus $(A, B, C)_s$ and $(\hat{A}, \hat{B}, \hat{C})_s$ are sign-similar, with T as above and scaling matrix $\Lambda = \Lambda_1^{-1} \Lambda_2$. \square

The rank of $\mathcal{A} = \{1, -1, 1, -1, \dots\}$ from Example 4.3, is 1, which is clearly finite. The triple $A = -1$, $B = 1$, $C = 1$ is a realization of the i/o map α , which is

controllable and observable; hence it is minimal. An example of a nonminimal sign-linear realization of the same α is

$$A = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}; B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}; C = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

In this case (A, B) is a controllable pair, but (A, C) is not an observable pair.

4.3. Counterexamples. Note that the converses of parts 1 and 2 of Theorem 4.16 are not true for discrete-time systems. If a sign-linear system is minimal, it is not necessarily observable. For example, the system with $x^+ = u$ and $y = \text{sign}(x)$ is minimal, but $A = 0$ so it is not observable. Also, a system may be final-state observable, and yet not be minimal. For example,

$$\begin{aligned} x^+ &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} x + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u, \\ y &= \text{sign} \left[\begin{pmatrix} 0 & 1 \end{pmatrix} x \right] \end{aligned}$$

is final-state observable. After k steps (for any $k \geq 1$) any state (x_1, x_2) ends up at (u_k, x_2) and x_2 can be identified. However, (A, C) is not an observable pair. If this system would be minimal, then the corresponding triple would be canonical by Lemma 4.11. But then (A, C) would have to be an observable pair. The minimal system for this i/o map is one of dimension 1, namely, $x^+ = x + u$, $y = \text{sign}(x)$.

5. Canonical realizations of sign-linear i/o maps. We noted that for sign-linear systems (unlike for linear systems) it is not true that a system is minimal if and only if it is canonical. The problem is that a minimal *discrete-time* sign-linear system may have $\det A = 0$, in which case it is not observable (Theorem 3.4). We may then ask—what is the canonical realization of a minimal sign-linear system which is guaranteed to exist by abstract realization theory ([12, §5.8])? The answer, for $p = 1$, is that for any α realizable by a sign-linear system, there exists a canonical (reachable and observable) system $\tilde{\Sigma}$ that realizes α , where $\tilde{\Sigma}$ is in the form of a cascade of a sign-linear system and shift registers. (In the general case, $p > 1$, the result has to be modified: we can only conclude that there is a system of this cascade form in which the minimal system may be embedded.)

We know there exists some canonical realization. We need only to show that there is a canonical realization of the form described above. Next we sketch the construction for the single-output case ($p = 1$). First find a minimal sign-linear realization Σ of α . Then we know (A, B, C) is a canonical triple and satisfies property \mathcal{P} (Lemmas 4.11 and 4.14). Perform a change of variables in the state space so that A has the form

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

with A_1 an $n_1 \times n_1$ invertible matrix and A_2 an $n_2 \times n_2$ nilpotent matrix. (Note that if Σ is already observable, then A is invertible and there is no A_2 . This Σ is already in the canonical form we are looking for.) Now the system equations have the form

$$\begin{aligned} x_1^+ &= A_1 x_1 + B_1 u, \\ x_2^+ &= A_2 x_2 + B_2 u, \\ y &= \text{sign}(C_1 x_1 + C_2 x_2). \end{aligned}$$

From now on, assume Σ has the form described above. Let κ be the relative degree of the system and $l := \min\{\kappa, n_2\}$. Let $\tilde{\Sigma}$ be the discrete-time system with state space $\mathbb{R}^{n-l} \times \{-1, 0, 1\}^l$, and system equations

$$(10) \quad \begin{aligned} \xi^+ &= F\xi + Gu, \\ \zeta_1^+ &= \text{sign}(\xi_{n-l} + \mathcal{A}_l u), \\ \zeta_2^+ &= \zeta_1, \\ &\vdots \\ \zeta_l^+ &= \zeta_{l-1}, \\ \eta &= \zeta_l, \end{aligned}$$

where $(F, G) = (A_1, B_1)$ when $l = n_2$, and when $l < n_2$,

$$F = \begin{pmatrix} A_1 & 0 & 0 \\ C_1 A_1^{n_2} & 0 & 0 \\ 0 & I & 0 \end{pmatrix}, \quad G = \begin{pmatrix} B_1 \\ \mathcal{A}_{n_2} \\ \mathcal{A}_{n_2-1} \\ \vdots \\ \mathcal{A}_{l+1} \end{pmatrix},$$

and I is the identity matrix of size $n_2 - l - 1$. (When $l = n_2 - 1$, there is no “ I ” part.) This can be seen as a cascade of a sign-linear system and shift registers.

LEMMA 5.1. *The system $\tilde{\Sigma}$ is the observable reduction of Σ .*

Proof. First we show that two states x and z are indistinguishable for Σ if and only if

$$(11) \quad x_1 = z_1$$

$$(12) \quad \begin{cases} CA^{n_2-1}x &= CA^{n_2-1}z \\ &\vdots \\ CA^{l+1}x &= CA^{l+1}z \\ CA^l x &= CA^l z \end{cases}$$

$$(13) \quad \begin{cases} \text{sign}(CA^{l-1}x) &= \text{sign}(CA^{l-1}z) \\ &\vdots \\ \text{sign}(Cx) &= \text{sign}(Cz). \end{cases}$$

In the case $l = n_2$, we have only (11) and (13). Suppose all equalities hold. Since $l <$ relative degree, the first l output terms for Σ are independent of the control. Then the last l equalities imply that the first l output terms coincide for x and z , for any input. Equations (12) imply that actually the first n_2 output terms coincide for x and z .

The remaining outputs only involve the first n_1 components of the state because of the nilpotency of A_2 . So if $x_1 = z_1$, then we see that all the remaining output terms are equal for initial states x and z . Thus, x and z are indistinguishable.

On the other hand, if x, z are indistinguishable, then using any control sequence, the outputs for the two initial states are always equal. In particular, the first l output terms are independent of the control so we obtain the last l equalities directly. For equalities (12) (in the case $l < n_2$) look at the next $n_2 - l$ output terms. If

$$CA^k x \neq CA^k z, \text{ for some } l \leq k \leq n_2 - 1,$$

then property \mathcal{P} and Remark 3.3 would imply that there is some control that would cause the k th output to be different for x, z , contradicting indistinguishability. Thus, those equalities hold too. Finally, for (11), we may focus on the output terms $y(k)$ for $k \geq n_2$. Indistinguishability implies that in particular, for the 0 control, all output terms are equal. Then $CA^k x_1 = CA^k z_1$ for all $k \geq n_2$. But (A_1, C_1) is an observable pair and $\det A_1 \neq 0$ so this implies $x_1 = z_1$.

Now consider the mapping $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^{n-l} \times \{-1, 0, 1\}^l$ given by $\phi(x) = (\xi, \zeta)$, where

$$\xi = \begin{pmatrix} x_1 \\ CA^{n_2-1}x \\ \vdots \\ CA^l x \end{pmatrix} \in \mathbb{R}^{n-l},$$

$$\zeta = \begin{pmatrix} \text{sign}(CA^{l-1}x) \\ \vdots \\ \text{sign}(CAx) \\ \text{sign}(Cx) \end{pmatrix} \in \{-1, 0, 1\}^l$$

and $\xi = x_1$ if $l = n_2$. We just proved that x and z are indistinguishable if and only if $\phi(x) = \phi(z)$. To show that the map is onto, we must show that for any $(\xi, \zeta) \in \mathbb{R}^{n-l} \times \{-1, 0, 1\}^l$, there is some $x \in \mathbb{R}^n$ so that $\phi(x) = (\xi, \zeta)$. Since (A, C) is an observable pair, (A_2, C_2) is also an observable pair. Thus, we may let x_1 be the first n_1 components of ξ and x_2 the solution to

$$\begin{pmatrix} C_2 A_2^{n_2-1} \\ \vdots \\ C_2 A_2 \\ C_2 \end{pmatrix} x_2 = \begin{pmatrix} \xi_{n_1+1} - C_1 A_1^{n_2-1} x_1 \\ \vdots \\ \xi_{n-l} - C_1 A_1^l x_1 \\ \zeta_1 - C_1 A_1^{l-1} x_1 \\ \vdots \\ \zeta_l - C_1 x_1 \end{pmatrix}.$$

Then clearly, $\phi(x) = (\xi, \zeta)$. Furthermore, it is easy to verify that ϕ commutes with the dynamics of Σ so it is a system morphism in the sense of [12], §5.8. \square

LEMMA 5.2. *The system $\tilde{\Sigma}$ is reachable and observable and realizes the same i/o behavior as Σ .*

Proof. Since $\tilde{\Sigma}$ is the observable reduction of Σ , and Σ is reachable, [12, Lemma 5.8.3] implies that $\tilde{\Sigma}$ is both reachable and observable with the same input/output behavior as Σ . \square

Example 5.3. Let Σ be the system with state space \mathbb{R}^2 and

$$\begin{aligned} x_1^+ &= u, \\ x_2^+ &= x_2 + u, \\ y &= \text{sign}(x_1 + x_2). \end{aligned}$$

Then Σ is minimal. But this sign-linear system is not observable, since $\det A = 0$. Perform a change of variables in the state space so that the A matrix is in the form discussed above. In the new coordinates, (z_1, z_2) , the equations take the form

$$z_1^+ = z_1 - u,$$

$$\begin{aligned} z_2^+ &= u, \\ y &= \text{sign}(-z_1 + z_2). \end{aligned}$$

The Markov sequence is $\mathcal{A} = \{2, 1, 1, 1, \dots\}$, $n_2 = 1$, relative degree = 1, so $l = 1$. The state space for $\tilde{\Sigma}$ is $\mathbb{R} \times \{-1, 0, 1\}$ and the equations for $\tilde{\Sigma}$ are

$$\begin{aligned} \xi^+ &= \xi - u, \\ \zeta^+ &= \text{sign}(\xi + 2u), \\ \eta &= \zeta. \end{aligned}$$

This system is reachable and observable.

6. Sampling. In this section we make some remarks about the time-sampling of sign-linear systems. This is the process of replacing a given continuous-time sign-linear system by the discrete-time one that results when only piecewise constant inputs (with a fixed sampling time) are used. The results in this section can be used to obtain the continuous-time results of Theorem 3.5 as a consequence of those of Theorem 3.4, and they clarify the differences between the two types of results, in particular, the fact that invertibility of the A matrix is not needed in the continuous-time case.

Remark 6.1. Suppose that (A, B, C) has property \mathcal{P} . Then the continuous-time sign-linear system $(A, B, C)_s$ is observable.

Proof. We will prove this by studying the associated sampled system. Using the notations and terminology in [12, §2.10], for each $\delta > 0$, the δ -sampled system corresponding to Σ is

$$\Sigma_\delta : \begin{cases} x^+ &= Fx + Gu, \\ y &= \text{sign}(Cx), \end{cases}$$

where $F = e^{\delta A}$, $G = A^{(\delta)}B$, and $A^{(\delta)} = \int_0^\delta e^{(\delta-s)A}ds$. We want to show that there is a $\delta > 0$ so that if (A, B, C) has property \mathcal{P} then the δ -sampled system satisfies condition 2 of Theorem 3.4. If this is true then the sampled system would be observable (clearly $\det e^{\delta A} \neq 0$). Hence, Σ is observable using only piecewise constant controls that are constant on intervals of length δ , and the result is proved.

Apply Kalman's sampling theorem (see [12, Prop. 5.2.11]), to the pair (A, \hat{C}) obtained by dropping the rows of C not in $I(\mathcal{A})$. For any δ satisfying

$$(14) \quad \delta(\lambda - \mu) \neq 2\pi ik, \quad k = \pm 1, \pm 2, \pm 3, \dots,$$

for every two eigenvalues λ, μ of A , we have that

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ q=0, \dots, n-1}} \ker(C_j(e^{\delta A})^q) = \{0\}.$$

What is left is to show that

$$I(\mathcal{A}) = I(\mathcal{A}_\delta),$$

where \mathcal{A}_δ is the Markov sequence of $(e^{\delta A}, A^{(\delta)}B, C)$. Note that $I(\mathcal{A}_\delta) \subseteq I(\mathcal{A})$ is always true for any δ , so the other inclusion is the interesting one. We will prove that if the k th row \mathcal{A}^k of \mathcal{A} is nonzero then the k th row \mathcal{A}_δ^k of \mathcal{A}_δ is nonzero for all δ satisfying (14). This will be done by showing the stronger result that \mathcal{A}^k and \mathcal{A}_δ^k have the same Hankel rank.

Fix a $k \in I(\mathcal{A})$. By restricting our attention to the linear system described by (A, B, C_k) , whose Markov sequence is \mathcal{A}^k and sampled-Markov sequence is \mathcal{A}_δ^k , we may, and will, assume without loss of generality that \mathcal{A} is a sequence with $p = 1$ and C has only one row. Thus we need to show that if \mathcal{A} is a Markov sequence with $p = 1$ represented by the triple (A, B, C) and if δ satisfies (14) then \mathcal{A}_δ , the Markov sequence of $(e^{\delta A}, A^{(\delta)}B, C)$, has the same Hankel rank as \mathcal{A} .

So let $\delta, \mathcal{A}, \mathcal{A}_\delta$ and (A, B, C) be as described. Next define a sequence $\mathcal{A}^{(\delta)}$ as follows. If

$$K(t) = \sum_{i=1}^{\infty} \mathcal{A}_i \frac{t^{i-1}}{(i-1)!},$$

then the output function for $\Sigma = (A, B, C)$ is $y(t) = \int_0^t K(t-s)u(s)ds$. If we restrict to sampled controls of length δ , then

$$y(l\delta) = \sum_{j=0}^{l-1} \left[\int_{j\delta}^{(j+1)\delta} K(l\delta - s)ds \right] u_{j+1}.$$

Letting

$$\mathcal{A}_j^{(\delta)} = \int_{j\delta}^{(j+1)\delta} K(l\delta - s)ds, \quad j = 0, 1, 2, \dots$$

we get

$$y(l\delta) = \sum_{j=0}^{l-1} \mathcal{A}_j^{(\delta)} u_{j+1}.$$

Look at any linear representation of the Markov sequence \mathcal{A} . Take the δ -sampled system for that representation. The Markov sequence for the δ -sampled system is $\mathcal{A}^{(\delta)}$. In particular, applied to the given triple (A, B, C) , this means that

$$\mathcal{A}^{(\delta)} = \mathcal{A}_\delta.$$

Take now a canonical representation (A^c, B^c, C^c) of \mathcal{A} of dimension n^c . Its eigenvalues, i.e., the eigenvalues of the matrix A^c , are among the eigenvalues of the (possibly non-canonical) original triple (A, B, C) . Thus, δ also satisfies $\delta(\lambda - \mu) \neq 2\pi ik$, $k = \pm 1, \pm 2, \pm 3, \dots$, for any two eigenvalues λ and μ of A^c . Then controllability and observability of (A^c, B^c, C^c) are preserved by sampling at this δ ; and thus the sampled triple $(e^{\delta A^c}, (A^c)^{(\delta)}B^c, C^c)$ is itself canonical and is a linear representation of $\mathcal{A}_\delta = \mathcal{A}^{(\delta)}$. The rank of a Markov sequence is equal to the dimension of a canonical linear representation of that sequence ([12, Cor. 5.5.7]). Therefore,

$$\text{rank } \mathcal{A}^{(\delta)} = n^c = \text{rank } \mathcal{A},$$

as desired. \square

REFERENCES

- [1] A.D. BACK AND A.C. TSOI, *FIR and IIR synapses, a new neural network architecture for time-series modeling*, Neural Computation, 3 (1991), pp. 375–385.

- [2] A.M. BAKSHO, S. DASGUPTA, J.S. GARNETT, AND C.R. JOHNSON, *On the similarity of conditions for an open-eye channel and for signed filtered error adaptive filter stability*, Proc. IEEE Conf. Decision and Control, Brighton, UK, Dec. 1991, IEEE Publications, 1991, pp. 1786–1787.
- [3] D. F. DELCHAMPS, *Extracting State Information from a Quantized Output Record*, Systems Control Lett., 13 (1989), pp. 365–372.
- [4] ———, *Controlling the Flow of Information in Feedback Systems with Measurement Quantization*, Proc. IEEE Conf. Decision and Control, Tampa, Dec. 1989, IEEE Publications, 1989, pp. 2355–2360.
- [5] ———, *Stabilizing a Linear System With Quantized State Feedback*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 916–924.
- [6] R.O. DUDA AND P.E. HART, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.
- [7] C.E. GILES, G.Z. SUN, H.H. CHEN, Y.C. LEE, AND D. CHEN, *Higher order networks recurrent and grammatical inference*, Advances in Neural Information Processing Systems 2, D.S. Touretzky, ed., Morgan Kaufmann, San Mateo, CA, 1990.
- [8] G.W. PULFORD, R.A. KENNEDY, AND B.D.O. ANDERSON, *Neural network structure for emulating decision feedback equalizers*, Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Toronto, Canada, May 1991, pp. 1517–1520.
- [9] P. RAMADGE, *On the Periodicity of Symbolic Observations of Piecewise Smooth Discrete-Time Systems*, IEEE Trans. Automat. Control, AC-35 (1990) pp. 807–813.
- [10] R. SCHWARZSCHILD (KOPLON) AND E.D. SONTAG, *Linear systems with constrained observations, Part I*, Report SYCON-91-07, Rutgers Center for Systems and Control, Rutgers University, May 1991.
- [11] H. SIEGELMANN AND E.D. SONTAG, *Turing computability with neural nets*, Appl. Math. Lett., 4 (1991) pp. 77–80.
- [12] E.D. SONTAG, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.
- [13] E.D. SONTAG AND H. SUSSMANN, *Backpropagation separates where perceptrons do*, Neural Networks, 4 (1991) pp. 243–249.

REPRESENTATIONS OF SYMMETRIC LINEAR DYNAMICAL SYSTEMS*

FABIO FAGNANI[†] AND JAN WILLEMS[‡]

Abstract. The purpose of this paper is to study static symmetries in linear time-invariant differential dynamical systems. The main result is a representation theorem which brings the symmetry strongly into evidence. This result is then applied to a number of examples involving permutations and rotations. We close by proving a general result on the representation of compact groups on the ring of unimodular polynomial matrices.

Key words. linear systems, symmetry, representations, canonical forms, group representations, permutations, rotations

AMS subject classifications. 93A, 93B

1. Introduction. *Symmetry* is a very appealing concept in many scientific endeavors. It plays a major role particularly in physics and in chemistry (for example, in crystallography). It has also been extensively studied in the classical theory of dynamical systems. A salient result in this area is Noether's theorem showing the equivalence of symmetries and conservation laws in Hamiltonian dynamics.

Also, many control problems will exhibit symmetry. For example, it is of interest to ascertain if a platform suspended on four pivots with a 90° rotation symmetry can be adequately stabilized by a control mechanism that also has this symmetry. Many mechanical systems will have a rotation symmetry, and an analogous question occurs in this case. A classical control problem that can be viewed as a symmetry question is whether an optimal controller for a time-invariant system will itself be time-invariant.

Although some interesting work has been done on symmetry questions in control, it is not a standard problem area. Notable contributions are the papers by Hazewinkel and Martin [5], [6] and Martin [8] motivated by certain questions related to the stabilization of linear systems by means of symmetric feedback control laws. Other places in control where symmetry problems have been studied are [1], [4], [11]–[13]. These authors are mainly concerned with nonlinear systems.

The purpose of the present paper is a fundamental study of symmetry in the context of linear systems described by differential equations. We will mainly consider representation questions. In a later paper, we plan to apply these results to control problems. The mathematical formulation follows the setting proposed in [14]. In a sense, the paper is a sequel to [2], where an elegant representation result has been obtained for *time-reversible* systems (cf. Theorem 2). In the present paper, we will study *static* symmetries and apply the representation results obtained especially to systems that are invariant under permutations or under rotations.

In an essential way, the paper uses the theory of group representations, a rather abstract area of mathematics whose original motivation lies very much in various aspects of symmetry. For an introduction to the theory of group representations, refer to [9].

* Received by the editors June 10, 1991; accepted for publication (in revised form) April 29, 1992. This research was supported by a collaborative grant from the Netherlands Organization for the Advancement of Scientific Research (NWO) and the Consiglio Nazionale delle Ricerche (CNR) and the European Economic Community Science Program contract SCI-0433-C(A).

[†] Scuola Normale Superiore, Piazza dei Cavalieri, 56100 Pisa, Italy.

[‡] Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, the Netherlands.

We close this introduction with a few words about notation and nomenclature.

Throughout, \mathbb{K} will denote \mathbb{R} or \mathbb{C} . Some of our results will be rather different for \mathbb{R} and for \mathbb{C} . \mathbb{K}^n denotes the n -dimensional (column) vectors over \mathbb{K} , and $\mathbb{K}^{n_1 \times n_2}$ denotes the matrices over \mathbb{K} with n_1 rows and n_2 columns. We will always consider vectors as columns and occasionally write the n -column vector x in term of its components as $\text{col}(x_1, x_2, \dots, x_n)$. A composite matrix will be written as $M = [M_1; M_2]$, and so forth; $\text{diag}(m_1, m_2, \dots, m_n)$ denotes then $n \times n$ diagonal matrix with (i, i) th element m_i . A similar notation will be used for block diagonal matrices. The determinant of a matrix is denoted as \det .

Let $f : A \rightarrow B$. For $A' \subset A$, the restriction of f to A' will be denoted as $f|_{A'}$. The map that identifies an element of $A' \subset A$ with the same element in A will be called the canonical injection. \ker means kernel, and im means image. The set of infinitely differentiable maps from A to B will be denoted as $C^\infty(A; B)$.

The set of polynomial matrices over \mathbb{K} with n_1 rows and n_2 columns in the indeterminate s will be denoted by $\mathbb{K}^{n_1 \times n_2}[s]$; $\mathbb{K}^{\bullet \times n}[s]$ denotes the set of polynomial matrices with n columns and any (of course, finite) number of rows. An element $R \in \mathbb{K}^{n_1 \times n_2}[s]$ is said to be of full row rank if it contains a $n_1 \times n_1$ submatrix with determinant nonzero. We will denote the set of full row rank elements of $\mathbb{K}^{\bullet \times n}[s]$ by $\mathbb{K}_{fr}^{\bullet \times n}[s]$; of $\mathbb{K}_{fr}^{n_1 \times n_2}[zs]$ denotes the elements of $\mathbb{K}_{fr}^{\bullet \times n_2}[s]$ with n_1 rows.

Let \mathcal{R} be a ring with an identity. An element $U \in \mathcal{R}$ is said to be *unimodular* if there exists $U^{-1} \in \mathcal{R}$ such that $UU^{-1} = U^{-1}U$ is equal to the identity. The unimodular elements of \mathcal{R} clearly form a multiplicative group, called the group of units of \mathcal{R} . The set of $n \times n$ matrices over \mathcal{R} also forms a ring. Its group of units will be denoted by $GL(n, \mathcal{R})$. The following two examples will be very important to us throughout the paper.

1. $GL(n, \mathbb{K})$, the set of nonsingular elements of $\mathbb{K}^{n \times n}$;
2. $GL(n, \mathbb{K}[s])$, the set of unimodular $(n \times n)$ polynomial matrices. Thus $U \in \mathbb{K}^{n \times n}[s]$ belongs to $GL(n, \mathbb{K}[s])$ if and only if its determinant is nonzero and belongs to \mathbb{K} , i.e., if it is a nonzero constant.

The set of isomorphisms on the vector space V is denoted by $GL(V)$. Thus, by considering the matrix representation of elements of $GL(\mathbb{K}^n)$ with respect to the standard basis, $GL(\mathbb{K}^n) \cong GL(n, \mathbb{K})$. As such, we will not make a distinction between these two sets and use $GL(\mathbb{K}^n)$ even where it may be more natural to write $GL(n, \mathbb{K})$.

Let M be a set. A *parametrization* (P, π) of M consists of a set P and a surjective map $\pi : P \rightarrow M$. The set P is called the parameter space. Typically, M is an abstract set, while P consists of concrete objects (as matrices or polynomial matrices—in which case, we refer to a matrix parametrization or a polynomial matrix parametrization of M). Note that π is surjective but not necessarily bijective. If π is a bijection, we will call the parametrization trim. In any case, the map $\pi : P \rightarrow M$ leads to the equivalence relation E on P defined by $(p_1 E p_2) :\Leftrightarrow (\pi(p_1) = \pi(p_2))$. This equivalence relation leads to canonical forms and to invariants. A subset $P_c \subseteq P$ will be called a *canonical form* for the parametrization (P, π) if $\pi(P_c) = M$, i.e., if $(P_c, \pi|_{P_c})$ is also parametrization of M . It is a trim canonical form if $\pi|_{P_c} : P_c \rightarrow M$ is a bijection.

2. Differential dynamical systems. Following the terminology explained in [15], we will define a *dynamical system* Σ to consist of a triple, $\Sigma = (\mathbb{T}, \mathbb{W}, \mathcal{B})$, with \mathbb{T} a subset of \mathbb{R} , called the *time axis*; \mathbb{W} a set called the *signal space*; and \mathcal{B} a subset of $\mathbb{W}^{\mathbb{T}}$ ($:=$ all maps from \mathbb{T} to \mathbb{W}), called the *behavior*. Thus the behavior consists of a given family of trajectories $w : \mathbb{T} \rightarrow \mathbb{W}$.

We will consider continuous-time dynamical systems with time axis $\mathbb{T} = \mathbb{R}$ and

with signal space $\mathbb{W} = \mathbb{K}^q$, with $\mathbb{K} = \mathbb{R}$ (the real case) or $\mathbb{K} = \mathbb{C}$ (the complex case). We will treat both cases in parallel. As we will see, there are distinct advantages, stemming from the theory of group representations, not to limit attention to the real case, although, admittedly, it is the real case that is of interest in applications. However, in §7.4 we see that rotation symmetries actually lead to complex systems!

The dynamical system $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B})$ is said to be *linear* if \mathcal{B} is a linear subspace of $(\mathbb{K}^q)^{\mathbb{R}}$ (the set of all maps from \mathbb{R} to \mathbb{K}^q) and *time-invariant* if $\sigma^t \mathcal{B} = \mathcal{B}$ for all $t \in \mathbb{R}$; σ^t denotes the backward t -shift (specifically, for $f : \mathbb{R} \rightarrow \mathbb{K}^q$ and $t \in \mathbb{R}$, $\sigma^t f : \mathbb{R} \rightarrow \mathbb{K}^q$ is defined by $(\sigma^t f)(t') := f(t + t')$).

In the present paper, we study behaviors \mathcal{B} that are the solution set of a system of constant coefficient linear differential equations

$$(1) \quad R \left(\frac{d}{dt} \right) w = 0$$

defined in terms of a polynomial matrix $R \in \mathbb{K}^{\bullet \times q}[s]$. The solution set of (1) is formally defined as follows:

$$\mathcal{B} = \left\{ w \in C^\infty(\mathbb{R}; \mathbb{K}^q) \mid \left(R \left(\frac{d}{dt} \right) w \right) (t) = 0 \text{ for all } t \in \mathbb{R} \right\}.$$

The assumption that w is infinitely differentiable is used mainly for convenience. The results may be generalized without difficulty to the case that \mathcal{B} also allows locally integrable functions, or distributions. However, for the purposes of the present paper, the smoothness assumption simplifies the analysis somewhat. In other applications, the C^∞ assumption may be very awkward.

The class of dynamical systems studied in this paper consists of those whose behavior is the kernel of a constant coefficient linear differential operator (with, for $R \in \mathbb{K}^{p \times q}[s]$, $R(d/dt)$ viewed as a map from $C^\infty(\mathbb{R}; \mathbb{K}^q)$ to $C^\infty(\mathbb{R}; \mathbb{K}^p)$). We will denote this class of dynamical systems as \mathcal{L}^q and refer to its elements as *differential dynamical systems*.

The above shows that $(\mathbb{K}^{\bullet \times q}[s], \pi)$ is a parametrization of \mathcal{L}^q with for $R \in \mathbb{K}^{\bullet \times q}[s]$, $\pi(R) := (\mathbb{R}, \mathbb{K}^q, \ker R(d/dt))$. This induces the equivalence relation \sim on $\mathbb{K}^{\bullet \times q}[s]$ defined by $(R_1 \sim R_2) \Leftrightarrow (\pi(R_1) = \pi(R_2))$. Note, in fact, that π is not injective. Indeed, if $R \in \mathbb{K}^{p \times q}[s]$ and $U \in GL(p, \mathbb{K}[s])$ (thus U is unimodular), then clearly $UR \sim R$.

We will call the system of differential equations (1) or, equivalently, R , a *behavioral equation representation* of $\pi(R)$; (1) or, equivalently, R is called a *minimal* (behavioral equation) representation of $\pi(R)$ if $(R_1 \in \mathbb{K}^{p_1 \times q}[s], R \in \mathbb{K}^{p \times q}[s], \text{ and } R_1 \sim R)$ implies $(p_1 \geq p)$. Let $\Sigma \in \mathcal{L}^q$ and let R be a minimal behavioral equation of Σ . Obviously, the number of rows of $R \in \mathbb{K}^{\bullet \times q}[s]$ will depend only on Σ but not on the particular minimal representation R of Σ . We will denote the number of rows of R by $p(\Sigma)$. Actually, $p(\Sigma)$ is equal to the number of output variables in any input/output representation of Σ (see [10]).

The following characterization of minimal representations will play an important role throughout the paper.

PROPOSITION 1. (1) *is minimal if and only if $R \in \mathbb{K}_{f_r}^{\bullet \times q}[s]$ (that is, $R \in \mathbb{K}^{\bullet \times q}[s]$ is of full row rank). Moreover, if (1) is minimal and if $R_1 \in \mathbb{K}^{\bullet \times q}[s]$, then $(R_1 \sim R \text{ and } R_1 \text{ is also minimal}) \Leftrightarrow (R_1 \text{ and } R \text{ both belong to } \mathbb{K}^{p(\Sigma) \times q}[s], \text{ and there exists a } U \in GL(p(\Sigma), \mathbb{K}[s]) \text{ such that } R_1 = UR)$. Finally, this U is unique.*

Proof. For the proof, see [10]. \square

This proposition implies that all minimal representations may be obtained from one by acting (as premultiplication) with the unimodular group. The freedom that this implies on the representations of a given dynamical system in terms of (minimal) behavioral equations will allow us to choose R 's in (1) that have an appealing form, reflecting symmetries.

3. Symmetric systems. The purpose of this paper is to study symmetries of dynamical systems in \mathcal{L}^q . A symmetry is induced by a transformation group, the basic idea being that we have a group of transformations mapping a dynamical system $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$ into another such dynamical system. If this transformation does not change Σ , then we will call Σ symmetric. We will now formalize this.

3.1. Transformation groups. Let S be a set and G be a group. Let T be a map from G into the group of bijections on S . We will denote the T -image of $g \in G$ by T_g . Then T is said to be a *transformation group* on S if T is a group homomorphism, that is, if $T_{g_1 g_2} = T_{g_1} T_{g_2}$. (The multiplication $g_1 g_2$ refers to the multiplication in the group G , while $T_{g_1} T_{g_2}$ refers to composition of maps on S .) For $s \in S$, the set $O_s := \{s' \in S \mid \exists g \in G \text{ such that } s' = T_g s\}$ is called the *orbit* through s . It is easily seen that, for $s_1, s_2 \in S$, either $O_{s_1} = O_{s_2}$, or $O_{s_1} \cap O_{s_2} = \emptyset$, the first situation occurring if and only if $s_2 \in O_{s_1}$. The collection of orbits $\{O_s \mid s \in S\}$ hence defines a partition of S and thus an equivalence relation on S .

3.2. Symmetries. Let T be a transformation group on \mathcal{L}^q . We will call the dynamical system $\Sigma \in \mathcal{L}^q$, *T-symmetric* if $T_g \Sigma = \Sigma$ for all $g \in G$. Thus, for a symmetric element Σ , the orbit O_Σ is equal to the singleton $\{\Sigma\}$.

Let us now consider a few examples of symmetries on \mathcal{L}^q .

Example 1 (time-invariance). Let $G = \mathbb{R}$ and define, for $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$, $T_g \Sigma$ as $T_g \Sigma = (\mathbb{R}, \mathbb{K}^q, \sigma^g \mathcal{B})$ (with σ^g the backward g -shift). It is easy to see that $T_g \Sigma = \Sigma$ for all $g \in \mathbb{R}$, and hence all elements of \mathcal{L}^q are symmetric in this sense. It is this symmetry that we call *time-invariance*. It formalizes the fact that the laws governing a dynamical system do not depend explicitly on time.

Example 2 (time-reversibility). Many examples of symmetries involve the group consisting of only two elements, $G = \{1, g\}$, $1 \neq g = g^{-1}$. Then $T_g = (T_g)^{-1}$; i.e. T_g is an *involution*. Define, for $\Sigma \in \mathcal{L}^q$, $T_g \Sigma$ as $T_g \Sigma := (\mathbb{R}, \mathbb{R}^q, \text{rev } \mathcal{B})$ with for $w : \mathbb{R} \rightarrow \mathbb{R}^q$, $\text{rev } w : \mathbb{R} \rightarrow \mathbb{R}^q$, the *time-reverse* of defined by $(\text{rev } w)(t) := w(-t)$. This Σ will be symmetric with respect to this transformation group if and only if $\mathcal{B} = \text{rev } \mathcal{B}$. This symmetry is called *time-reversibility*. It expresses the fact that the system looks identical when viewed backward in time. We have studied this symmetry in detail in [2] and will return to it later in this paper.

3.3. Static symmetries. Let T be a transformation group acting on \mathbb{K}^q ; T induces a symmetry on \mathcal{L}^q by defining for $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$, $T_g \Sigma$ as $T_g \Sigma := (\mathbb{R}, \mathbb{K}^q, T_g \mathcal{B})$ with $T_g \mathcal{B} := \{w : \mathbb{R} \rightarrow \mathbb{R}^q \mid \exists w' : \mathbb{R} \rightarrow \mathbb{K}^q \text{ such that } w(t) = T_g w'(t) \text{ for all } t \in \mathbb{R}\}$. Note that, by a minor abuse of notation, we use the same symbol T_g as acting on \mathcal{L}^q , on \mathcal{B} , and on \mathbb{K}^q . Thus Σ is symmetric in this sense if $w \in \mathcal{B}$ implies $T_g w \in \mathcal{B}$ for all $g \in G$. Since T_g transforms the trajectories w in \mathcal{B} by applying the memoryless map T_g (that is, since it transforms trajectories w in a nondynamic way), we will call such a symmetry a *static symmetry*. In fact, we will be particularly interested in the case where T_g is linear for all $g \in G$. Such transformation groups are the subject of the theory of group representations. It is customary to denote T by ρ in that case.

3.4. Group representations. Let V be a vector space over the field \mathbb{K} . A group homomorphism $\rho : G \rightarrow GL(V)$ is said to be a (linear) *representation* of the group G on V . If V is finite-dimensional, then the representation is called *finite-dimensional*, and the dimension of V as a vector space is called the *order* of the representation. In particular, if V is n -dimensional and if we represent elements of $GL(V)$ as matrices with respect to a fixed basis on V , then a representation of G will correspond to each element of the group G , a nonsingular $(n \times n)$ matrix over \mathbb{K} such that group multiplication goes over in multiplication of matrices. In particular, the identity matrix will correspond to the unit element in G .

Example 3 (permutations). As a specific example of a static symmetry, let S_q denote the group of permutations of q elements. This group is called the *symmetric group*; it is a finite group consisting of $q!$ elements. Now consider the map $\rho : S_q \rightarrow GL(\mathbb{K}^q)$, which associates with the permutation $g : \{1, 2, \dots, q\} \rightarrow \{1, 2, \dots, q\}$ the linear bijection on \mathbb{K}^q that takes the vector $\text{col}(x_1, x_2, \dots, x_q)$ into the vector $\text{col}(x_{g(1)}, x_{g(2)}, \dots, x_{g(q)})$. Clearly, ρ defines a representation of S_q on \mathbb{K}^q . Thus ρ in this case maps onto the group of $q \times q$ permutation matrices. A dynamical system $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B})$ will be symmetric in the sense of the static symmetry induced by this representation of S_q , provided that $w = \text{col}(w_1, w_2, \dots, w_q) \in \mathcal{B}$ if and only if $w' = \text{col}(w'_1, w'_2, \dots, w'_q) \in \mathcal{B}$ with $(w'_1, w'_2, \dots, w'_q)$ any permutation of (w_1, w_2, \dots, w_q) . We can think of this symmetry as occurring when Σ models the dynamics of the positions of q identical particles on the line: feasible motions will remain feasible motions after we interchange the positions of the particles. More meaningful symmetries as representations of S_q (involving particles in the plane or in 3-space), or of subgroups of S_q , will be considered later.

Let $\rho : G \rightarrow GL(V)$ be a representation of G on V , with V a finite-dimensional vector space over \mathbb{K} . Throughout this paper, we will assume that G is either finite or compact. In the compact case, G is assumed to be a compact Hausdorff topological space with the group multiplication and the inverse continuous maps. A representation $\rho : G \rightarrow GL(V)$ is then always assumed to be continuous.

A subspace $V_1 \subseteq V$ is said to be *invariant* if $\rho_g V_1 \subseteq V_1$ for all $g \in G$. The representation ρ is said to be *irreducible* if its only invariant subspaces are V and $\{0\}$. When V_1 is invariant, then ρ^{V_1} , defined by $\rho^{V_1} : G \rightarrow GL(V_1)$ with $\rho^{V_1}_g := \rho_g|_{V_1}$, yields another finite-dimensional representation of G : ρ^{V_1} is called a subrepresentation. It is a standard result from the theory of group representations that, if G is compact, then a finite-dimensional subrepresentation can be written as the direct sum of irreducible representations.

Let $\rho^1 : G \rightarrow GL(V_1)$ and $\rho^2 : G \rightarrow GL(V_2)$ be two finite-dimensional representations of the same group G . Then they are said to be *isomorphic* if V_1 and V_2 have the same dimension and if there exists an isomorphism $S : V_1 \rightarrow V_2$ such that $\rho^2_g = S \rho^1_g S^{-1}$ for all $g \in G$. Isomorphism of ρ_1 and ρ_2 will be denoted by $\rho_1 \cong \rho_2$. If ρ_1 is not isomorphic to ρ_2 , then ρ_1 and ρ_2 are said to be *distinct*.

Thus the above implies that a representation ρ admits a decomposition of the following type:

$$\rho \cong m_1 \rho_1 \oplus m_2 \rho_2 \oplus \cdots \oplus m_k \rho_k,$$

where $\rho_1 \cdots \rho_k$ are distinct irreducible representation and where

$$m_j \rho_j := \underbrace{\rho_j \oplus \rho_j \oplus \cdots \oplus \rho_j}_{m_j \text{--times}}.$$

Example 3 (continued). Recall that S_q denotes the group of permutations of q elements. It is a finite group containing $q!$ elements. The irreducible representations of S_q have been studied in much detail in the literature. However, we will need only two of them. Consider the following representations of S_q :

1. The *identity representation*, $\rho_1 : S_q \rightarrow GL(\mathbb{K})$ with $\rho_{1,g} = 1$ for all $g \in S_q$. This representation is of order 1 and hence irreducible;

2. The representation ρ_2 defined as follows. Let V be the subspace of \mathbb{K}^q consisting of those vectors $\text{col}(x_1, x_2, \dots, x_q)$ such that $\sum_{i=1}^q x_i = 0$. Let S_q act on V by $\rho_{2,g} \text{col}(x_1, x_2, \dots, x_q) := \text{col}(x_{g(1)}, x_{g(2)}, \dots, x_{g(q)})$. It is easy to prove that ρ_2 defines an irreducible representation of S_q . Since $\dim V = q - 1$, its order is $q - 1$.

Let $\rho : S_q \rightarrow GL(\mathbb{K}^q)$ be the representation of S_q introduced in Example 3: $\rho_g \text{col}(x_1, x_2, \dots, x_q) := \text{col}(x_{g(1)}, x_{g(2)}, \dots, x_{g(q)})$. Write $\mathbb{K}^q = V_1 \oplus V_2$ with $V_1 = \{\text{col}(x_1, x_2, \dots, x_q) \in \mathbb{K}^q \mid x_1 = x_2 = \dots = x_q\}$ and $V_2 = \{\text{col}(x_1, x_2, \dots, x_q) \in \mathbb{K}^q \mid x_1 + x_2 + \dots + x_q = 0\}$. Clearly, V_1 and V_2 are ρ -invariant subspaces, $\rho^{V_1} \cong \rho_1$ and $\rho^{V_2} \cong \rho_2$. Hence, in this case, the decomposition of ρ in terms of the irreducible representations of S_q becomes $\rho \cong \rho_1 \oplus \rho_2$. (We hence have $m_1 = m_2 = 1$, while all the other m_i 's are zero. Furthermore, $n_1 = 1$ and $n_2 = q - 1$.)

4. Representation questions for symmetric systems. Assume that $\rho : G \rightarrow GL(\mathbb{K}^q)$ is a representation of the group G on \mathbb{K}^q and assume that $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$ is symmetric in the sense of the static symmetry induced by this representation. The problem studied in this paper is the following: *Can this symmetry be put into evidence by an appropriate behavioral equation representation of Σ as (1), in which the polynomial matrix R is such that this static symmetry becomes evident?* Otherwise stated, we want to come up with a *parametrization*, with a *canonical form* for the behavioral equations of systems with a static symmetry.

To give an example of the type of results that we seek, we repeat the main result of [2]. This result involves time-reversibility, which, it should be noted, is not a static symmetry.

THEOREM 2. $\Sigma \in \mathcal{L}^q$ is time-reversible if and only if it allows a minimal behavioral equation representation (1) with $R(s) = JR(-s)$ with J a matrix of the type

$$J = \begin{bmatrix} I_1 & 0 \\ 0 & -I_2 \end{bmatrix},$$

where I_1 and I_2 are identity matrices.

Observe that $R(s) = JR(-s)$ is equivalent to stating that (1) consists of a number of scalar differential equations, some of which contain only even-order derivatives, while the others contain only odd-order derivatives. If the equations in (1) are indeed of this form, then time-reversibility is obvious. Note, in particular, that time-reversible systems cannot always be represented by differential equations containing only even-order derivatives. (Actually, any representation as obtained in Theorem 2 will have the dimension of I_1 and I_2 as invariants.)

While this paper is only concerned with static symmetries, it is worthwhile noting that it is possible to view also reciprocity, a much-studied property of electrical networks [17], as a (dynamic) symmetry.

5. The main result. Assume in this section that $\rho : G \rightarrow GL(\mathbb{K}^q)$ is a given representation of a compact group G on \mathbb{K}^q . Then ρ defines a static symmetry on \mathcal{L}^q as described in §3.4; $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$ is thus ρ -symmetric if and only if $\rho_g \mathcal{B} = \mathcal{B}$ for all $g \in G$.

Let (1) be a minimal representation for such a ρ -symmetric $\Sigma \in \mathcal{L}^q$. It then follows immediately from Proposition 1 that, for each $g \in G$, there will exist a unimodular polynomial matrix $U_g(s)$ such that $R(s)\rho_g = U_g(s)R(s)$. Our main result tells us that R can be chosen such that $U_g(s)$ is a constant nonsingular matrix, thus independent of s !

THEOREM 3. *Let $\rho : G \rightarrow GL(\mathbb{K}^q)$ be a representation of the compact group G on \mathbb{K}^q . $\Sigma \in \mathcal{L}^q$ is ρ -symmetric if and only if there exists a minimal representation $R(d/dt)w = 0$ of Σ and a representation $\rho' : G \rightarrow GL(\mathbb{K}^{p(\Sigma)})$ of the group G on $\mathbb{K}^{p(\Sigma)}$ such that*

$$R(s)\rho_g = \rho'_g R(s)$$

for all $g \in G$. Moreover, ρ' will be isomorphic to a subrepresentation of ρ .

Proof. To prove the “if” part, assume that $R(s)\rho_g = \rho'_g R(s)$. Then, since ρ'_g is an invertible matrix (hence a unimodular polynomial matrix), $\ker R(d/dt) = \ker R(d/dt)\rho_g$. Hence $\rho_g \mathcal{B} = \mathcal{B}$ for all $g \in G$; ρ -symmetry follows. As a general feature of the type of representation results that we seek, note that also here (as in Theorem 2) the “if” is immediate: if $R(s)\rho_g = \rho'_g R(s)$ for all $g \in G$, then ρ -symmetry of (1) is basically immediate. The converse however is more difficult.

The “only if” part is based on Theorems 4 and 5 and will be proved later.

To see that ρ' is isomorphic to a subrepresentation of ρ , pick an element $\lambda \in \mathbb{R}$ such that $R(\lambda)$ has full row rank $p(\Sigma)$. Since R is minimal and hence of full row rank as a polynomial matrix, such a $\lambda \in \mathbb{K}$ exists. Now observe that $R(\lambda)\rho_g = \rho'_g R(\lambda)$ for all $g \in G$. Let $N := \ker R(\lambda)$. Obviously, N is ρ -invariant. Hence there exists a linear subspace M of \mathbb{K}^q such that M is ρ -invariant and $\mathbb{K}^q = N \oplus M$. Therefore $R(\lambda) \mid_M \rho_g \mid_M = \rho'_g R(\lambda) \mid_M$. Since $R(\lambda) \mid_M$ is a bijection, this shows that ρ' is isomorphic to the subrepresentation ρ^M of ρ . \square

6. Canonical forms for symmetric systems. We will now show that establishing Theorem 3 is equivalent to establishing the existence of a very nice explicit canonical forms for symmetric systems. At this point, it becomes necessary to treat the complex case ($\mathbb{K} = \mathbb{C}$) and the real case ($\mathbb{K} = \mathbb{R}$) separately.

6.1. Complex systems. The representations $\rho : G \rightarrow GL(\mathbb{C}^q)$ and $\rho' : G \rightarrow GL(\mathbb{C}^{p(\Sigma)})$ obtained in §5 can be decomposed in terms of irreducible ones as

$$\begin{aligned} \rho &\cong m_1 \rho_1 \oplus m_2 \rho_2 \oplus \cdots \oplus m_k \rho_k, \\ \rho' &\cong m'_1 \rho_1 \oplus m'_2 \rho_2 \oplus \cdots \oplus m'_k \rho_k. \end{aligned}$$

Since ρ' is isomorphic to a subrepresentation of ρ , it follows that the integers $m'_i \in \mathbb{Z}_+$ satisfy

$$0 \leq m'_i \leq m_i, \quad i = 1, 2, \dots, k$$

The above decomposition of ρ implies that there exists a nonsingular matrix $V \in \mathbb{C}^{q \times q}$ such that

$$V \rho_g V^{-1} = \text{diag} (m_1 \rho_{1,g}, \dots, m_k \rho_{k,g}) =: \tilde{\rho}_g,$$

where

$$m_i \rho_{i,g} := \text{diag} (\underbrace{\rho_{i,g}, \dots, \rho_{i,g}}_{m_i \text{--times}}).$$

Proceeding in a similar manner for ρ' , we obtain a nonsingular matrix $V' \in \mathbb{C}^{p(\Sigma) \times p(\Sigma)}$ such that

$$V' \rho'_g (V')^{-1} = \text{diag} (m'_1 \rho_{1,g}, m'_2 \rho_{2,g}, \dots, m'_k \rho_{k,g}) =: \tilde{\rho}'_g.$$

Note that applying the nonsingular transformation V corresponds to changing the signal variables in $\Sigma = (\mathbb{R}, \mathbb{C}^q, \mathcal{B})$ from $w : \mathbb{R} \rightarrow \mathbb{C}^q$ to $\tilde{w} : \mathbb{R} \rightarrow \mathbb{C}^q$ with $\tilde{w}(t) := Vw(t)$; in other words, it corresponds to choosing a convenient basis in the signal space \mathbb{C}^q . We will call such a basis a ρ -adapted basis, and the corresponding coordinates ρ -adapted coordinates (these are sometimes called *normal coordinates*). On the other hand, premultiplying R in (1) by nonsingular matrix V' corresponds to choosing a convenient basis in the equation space $\mathbb{C}^{p(\Sigma)}$. By Proposition 1, this does not change the behavior, and hence we can always assume that we are using such a basis on the equation space.

Now, assume that ρ and ρ' satisfy the conditions of Theorem 3. It follows thus that in a ρ -adapted basis the system $\Sigma = (\mathbb{R}, \mathbb{C}^q, \mathcal{B}) \in \mathcal{L}^q$ will admit a minimal representation

$$(2) \quad \tilde{R} \left(\frac{d}{dt} \right) \tilde{w} = 0,$$

where $\tilde{R} \in \mathbb{C}^{p(\Sigma) \times q}[s]$ satisfies

$$(3) \quad \tilde{R}(s) \tilde{\rho}_g = \tilde{\rho}'_g \tilde{R}(s).$$

Now partition \tilde{R} conformably as $\tilde{\rho}$ and $\tilde{\rho}'$, yielding

$$(4) \quad \tilde{R}(s) = \begin{bmatrix} \tilde{R}_{11}(s) & \tilde{R}_{12}(s) & \cdots & \tilde{R}_{1k}(s) \\ \tilde{R}_{21}(s) & \tilde{R}_{22}(s) & \cdots & \tilde{R}_{2k}(s) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{R}_{k1}(s) & \tilde{R}_{k2}(s) & \cdots & \tilde{R}_{kk}(s) \end{bmatrix}.$$

Then (3) implies that $\tilde{R}_{ij}(s) m_j \rho_j = m'_i \rho_i \tilde{R}_{ij}(s)$. By the Schur lemma [9], these equalities imply the following strong conclusions about \tilde{R} :

$$(5) \quad \tilde{R}_{ij} = 0 \quad \text{for } i \neq j$$

and

$$\tilde{R}_{ii}(s) = \begin{bmatrix} \lambda_{11}(s) I_{n_i} & \lambda_{12}(s) I_{n_i} & \cdots & \lambda_{1m_i}(s) I_{n_i} \\ \lambda_{21}(s) I_{n_i} & \lambda_{22}(s) I_{n_i} & \cdots & \lambda_{2m_i}(s) I_{n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m'_i 1}(s) I_{n_i} & \lambda_{m'_i 2}(s) I_{n_i} & \cdots & \lambda_{m'_i m_i}(s) I_{n_i} \end{bmatrix},$$

where n_i is the order of the representation ρ_i . In the Kronecker product notation, \tilde{R}_{ii} may be written as

$$(6) \quad \tilde{R}_{ii}(s) = \Lambda_i(s) \otimes I_{n_i},$$

where

$$\Lambda_i(s) = \begin{bmatrix} \lambda_{11}(s) & \lambda_{12}(s) & \cdots & \lambda_{1m_i}(s) \\ \lambda_{21}(s) & \lambda_{22}(s) & \cdots & \lambda_{2m_i}(s) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m'_i 1}(s) & \lambda_{m'_i 2}(s) & \cdots & \lambda_{m'_i m_i}(s) \end{bmatrix}.$$

This proves that, for $\mathbb{K} = \mathbb{C}$, Theorem 3 is equivalent to the following theorem.

THEOREM 4. *Let G be a compact group and let $\rho : G \rightarrow GL(\mathbb{C}^q)$ be a representation of G on \mathbb{C}^q . Assume that $\rho \cong m_1\rho_1 \oplus m_2\rho_2 \oplus \cdots \oplus m_k\rho_k$, with $\rho_i : G \rightarrow GL(\mathbb{C}^{n_i})$, $i = 1, 2, \dots, k$, distinct irreducible representations. Assume that the basis in \mathbb{C}^q is ρ -adapted (to emphasise this, we write the signal variables as \tilde{w} , $\tilde{w} = \text{col}(\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_k)$ with $\tilde{w}_i : \mathbb{R} \rightarrow (\mathbb{C}^{n_i})^{m_i}$). Then $\Sigma = (\mathbb{R}, \mathbb{C}^q, \mathcal{B}) \in \mathcal{L}^q$ is ρ -symmetric if and only if there exist $m'_i \in \mathbb{Z}_+$, $0 \leq m'_i \leq m_i$, and polynomial matrices $\Lambda_i \in \mathbb{C}_{fr}^{m'_i \times m_i}[s]$ such that Σ admits a minimal representation*

$$(7) \quad \left(\Lambda_i \left(\frac{d}{dt} \right) \otimes I_{n_i} \right) \tilde{w}_i = 0, \quad i = 1, 2, \dots, k.$$

Note that, from the above theorem, we may conclude that the ρ -adapted variables $\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_k$ are completely noninteracting!

6.2. Real systems. In the case ($\mathbb{K} = \mathbb{R}$) of systems with signal space \mathbb{R}^q , Theorem 4 remains, of course, valid but may yield a representation (7) with complex coefficients. However, in this case, we want to obtain differential equations in a canonical form analogous to (7) but with real coefficients. The theory becomes more involved, since the irreducible representations ρ_i introduced in §6.1 are irreducible over \mathbb{C} and need not be real. In particular, Schur's lemma in the form in which it was used in §6.1 would yield complex representations. Nevertheless, quite explicit results may also be obtained now.

As it is shown in [9], a irreducible representation $\rho : G \rightarrow GL(\mathbb{C}^n)$ can be of *real* type, of *complex* type, or of *quaternionic* type; ρ is of real type if it is the complexification of a real representation. Now, if ρ is an irreducible representation of complex or of quaternionic type, then so will be its complex conjugate, ρ^* (thus the matrices ρ_g and ρ_g^* are complex conjugates for all $g \in G$). By combining $\rho \oplus \rho^*$, these complex representations lead to real ones. All together, this leads to the following decomposition of a real representation.

A real representation $\rho : G \rightarrow GL(\mathbb{R}^q)$ of a compact group G admits a decomposition

$$(8a) \quad \rho \cong \rho_{\mathbb{R}} \oplus \rho_{\mathbb{C}} \oplus \rho_{\mathbb{H}}$$

with $\rho_{\mathbb{R}}$, $\rho_{\mathbb{C}}$, and $\rho_{\mathbb{H}}$ referring to the further decomposition into irreducible representations of real, complex, or quaternionic type. These representations can indeed be further decomposed as follows:

$$(8b) \quad \rho_{\mathbb{R}} = m_{\mathbb{R},1}\rho_{\mathbb{R},1} \oplus m_{\mathbb{R},2}\rho_{\mathbb{R},2} \oplus \cdots \oplus m_{\mathbb{R},k}\rho_{\mathbb{R},k},$$

$$(8c) \quad \rho_{\mathbb{C}} = m_{\mathbb{C},1}\rho_{\mathbb{C},1} \oplus m_{\mathbb{C},2}\rho_{\mathbb{C},2} \oplus \cdots \oplus m_{\mathbb{C},k}\rho_{\mathbb{C},k},$$

$$(8d) \quad \rho_{\mathbb{H}} = m_{\mathbb{H},1}\rho_{\mathbb{H},1} \oplus m_{\mathbb{H},2}\rho_{\mathbb{H},2} \oplus \cdots \oplus m_{\mathbb{H},k}\rho_{\mathbb{H},k}$$

with $\rho_{\mathbb{R},i} : G \rightarrow GL(\mathbb{R}^{n_{\mathbb{R},i}})$ such that its complexification is irreducible over \mathbb{C} ; $\rho_{\mathbb{C},i} : G \rightarrow GL(\mathbb{R}^{2n_{\mathbb{C},i}})$ such that the complexification of $\rho_{\mathbb{C},i}$ is isomorphic (over \mathbb{C}) to $\tilde{\rho}_{\mathbb{C},i} \oplus \tilde{\rho}_{\mathbb{C},i}^*$ with $\tilde{\rho}_{\mathbb{C},i} : G \rightarrow GL(\mathbb{C}^{n_{\mathbb{C},i}})$ irreducible over \mathbb{C} ; similarly, $\rho_{\mathbb{H},i} : G \rightarrow GL(\mathbb{R}^{4n_{\mathbb{H},i}})$, $\rho_{\mathbb{H},i} \cong \tilde{\rho}_{\mathbb{H},i} \oplus \tilde{\rho}_{\mathbb{H},i}^*$ with $\tilde{\rho}_{\mathbb{H},i} : G \rightarrow GL(\mathbb{C}^{2n_{\mathbb{H},i}})$ also irreducible over \mathbb{C} .

The representations $\rho_{\mathbb{R},1}, \dots, \rho_{\mathbb{R},k_{\mathbb{R}}}; \tilde{\rho}_{\mathbb{C},1}, \dots, \tilde{\rho}_{\mathbb{C},k_{\mathbb{C}}}; \tilde{\rho}_{\mathbb{C},1}^*, \dots, \tilde{\rho}_{\mathbb{C},k_{\mathbb{C}}}^*; \tilde{\rho}_{\mathbb{H},1}, \dots, \tilde{\rho}_{\mathbb{H},k_{\mathbb{H}}}$ are the distinct irreducible representations of real, complex, and quaternionic type, respectively, involved in the decomposition of ρ ,

$$n_{\mathbb{R},1}, \dots, n_{\mathbb{R},k_{\mathbb{R}}}; n_{\mathbb{C},1}, \dots, n_{\mathbb{C},k_{\mathbb{C}}}; n_{\mathbb{H},1}, \dots, n_{\mathbb{H},k_{\mathbb{H}}},$$

their respective orders, and

$$m_{\mathbb{R},1}, \dots, m_{\mathbb{R},k_{\mathbb{R}}}; m_{\mathbb{C},1}, \dots, m_{\mathbb{C},k_{\mathbb{C}}}; m_{\mathbb{H},1}, \dots, m_{\mathbb{H},k_{\mathbb{H}}},$$

their respective multiplicities.

A real subrepresentation ρ' of ρ will allow a similar representation as (8), but with the analogous multiplicities $m'_{\mathbb{R},i}$, $m'_{\mathbb{C},i}$ and $m'_{\mathbb{H},i}$ satisfying $0 \leq m'_{\mathbb{R},i} \leq m_{\mathbb{R},i}$, $0 \leq m'_{\mathbb{C},i} \leq m_{\mathbb{C},i}$ and $0 \leq m'_{\mathbb{H},i} \leq m_{\mathbb{H},i}$.

Now, assume that ρ and ρ' satisfy the conditions of Theorem 3. Proceeding as in the complex case, we obtain (in the analogous partition of \tilde{R}) that, in a basis compatible with (8), $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathcal{B}) \in \mathcal{L}^q$ will admit a (real) minimal representation as (2) with (5) still satisfied. Hence the off-diagonal blocks of \tilde{R} will still be zero: the components of \tilde{w} will again be noninteracting.

However, Schur's lemma allows us to conclude the simple form (6) for the diagonal blocks of \tilde{R} only for the diagonal blocks corresponding to the real representations, the $\rho_{\mathbb{R},i}$'s. The diagonal blocks of \tilde{R} corresponding to the $\rho_{\mathbb{C},i}$'s and the $\rho_{\mathbb{H},i}$'s will be more complicated, and it is here that the difference between the real, complex, and quaternionic type plays a role. To obtain a convenient form for the corresponding diagonal blocks of \tilde{R} , we should choose the basis in the $\mathbb{R}^{2n_{\mathbb{C},i}}$'s such that $\rho_{\mathbb{C},i}$ takes the form

$$(9a) \quad \rho_{\mathbb{C},i} \cong \begin{bmatrix} A_{\mathbb{C},i} & -B_{\mathbb{C},i} \\ B_{\mathbb{C},i} & A_{\mathbb{C},i} \end{bmatrix},$$

and in the $\mathbb{R}^{4n_{\mathbb{H},i}}$'s such that $\rho_{\mathbb{H},i}$ takes the form

$$(9b) \quad \rho_{\mathbb{H},i} \cong \begin{bmatrix} A_{\mathbb{H},i} & -B_{\mathbb{H},i} & -C_{\mathbb{H},i} & -D_{\mathbb{H},i} \\ B_{\mathbb{H},i} & A_{\mathbb{H},i} & D_{\mathbb{H},i} & -C_{\mathbb{H},i} \\ C_{\mathbb{H},i} & -D_{\mathbb{H},i} & A_{\mathbb{H},i} & B_{\mathbb{H},i} \\ D_{\mathbb{H},i} & C_{\mathbb{H},i} & -B_{\mathbb{H},i} & A_{\mathbb{H},i} \end{bmatrix}.$$

It can be shown that there exists a (real) choice of the basis in \mathbb{R}^q that is compatible with the decomposition (8a) and in which the $\rho_{\mathbb{C},i}$'s and the $\rho_{\mathbb{H},i}$'s have the above form. We will call such a basis choice *real ρ -adapted*.

Schur's lemma then allows to conclude that in a real ρ -adapted basis in \mathbb{R}^q and in a real ρ' -adapted basis in $\mathbb{R}^{p(\Sigma)}$ we will obtain a representation (2) with

$$(10a) \quad \tilde{R}_{ij}(s) = 0 \quad \text{for } i \neq j$$

and

$$(10b) \quad \tilde{R}_{ii}(s) = \tilde{A}_i(s) \otimes I_{n_{\mathbb{R},i}}$$

for the diagonal blocks corresponding to the $\rho_{\mathbb{R},i}$'s

$$(10c) \quad \tilde{R}_{ii}(s) = \begin{bmatrix} \tilde{A}_i(s) & -\tilde{B}_i(s) \\ \tilde{B}_i(s) & \tilde{A}_i(s) \end{bmatrix} \otimes I_{n_{\mathbb{C},i}}$$

for the diagonal blocks corresponding to the $\rho_{C,i}$'s, and

$$(10d) \quad \tilde{R}_{ii}(s) = \begin{bmatrix} \tilde{A}_i(s) & -\tilde{B}_i(s) & -\tilde{C}_i(s) & -\tilde{D}_i(s) \\ \tilde{B}_i(s) & \tilde{A}_i(s) & -\tilde{D}_i(s) & \tilde{C}_i(s) \\ \tilde{C}_i(s) & \tilde{D}_i(s) & \tilde{A}_i(s) & -\tilde{B}_i(s) \\ \tilde{D}_i(s) & -\tilde{C}_i(s) & \tilde{B}_i(s) & \tilde{A}_i(s) \end{bmatrix} \otimes I_{n_{H,i}}$$

for the diagonal blocks corresponding to the $\rho_{H,i}$'s.

The notation may be further streamlined by using complex numbers in (9a) and (10c) and quaternions in (9b) and (10d). Coding a typical vector $\text{col}(\tilde{w}_{C,i}^1, \tilde{w}_{C,i}^2) \in \mathbb{R}^{2n_{C,i}}$ in (9a) as the complex vector $\tilde{w}_{C,i}^1 + i\tilde{w}_{C,i}^2 \in \mathbb{C}^{n_{C,i}}$, ensures that multiplication by (10c) corresponds to multiplication by the complex polynomial matrix $(\tilde{A}_i(s) + i\tilde{B}_i(s)) \otimes I_{n_{C,i}}$. Coding a typical vector $\tilde{w}_{H,i} = \text{col}(\tilde{w}_{H,i}^1, \tilde{w}_{H,i}^2, \tilde{w}_{H,i}^3, \tilde{w}_{H,i}^4)$ in (9b) as the quaternionic vector $\tilde{w}_{H,i}^1 + i\tilde{w}_{H,i}^2 + j\tilde{w}_{H,i}^3 + k\tilde{w}_{H,i}^4$ ensures that multiplication by (10d) corresponds to multiplication by the quaternionic polynomial matrix $(\tilde{A}_i(s) + i\tilde{B}_i(s) + j\tilde{C}_i(s) + k\tilde{D}_i(s)) \otimes I_{n_{H,i}}$. For the multiplication rules for quaternions, see Example 8 and §10.

In the following theorem, we assume that the real ρ -adapted basis has been streamlined in this way. These considerations prove that, for $\mathbb{K} = \mathbb{R}$, Theorem 3 is equivalent to the following theorem.

THEOREM 5. *Let G be a compact group and let $\rho : G \rightarrow GL(\mathbb{R}^q)$ be a representation of G on \mathbb{R}^q . Assume that ρ is decomposed as (8) and that the basis in \mathbb{R}^q is real ρ -adapted. To emphasize this, we write the signal variables as \tilde{w} ,*

$$\begin{aligned} \tilde{w} &= \text{col}(\tilde{w}_{\mathbb{R}}, \tilde{w}_{\mathbb{C}}, \tilde{w}_{\mathbb{H}}), \\ \tilde{w}_{\mathbb{R}} &= \text{col}(\tilde{w}_{\mathbb{R},1}, \tilde{w}_{\mathbb{R},2}, \dots, \tilde{w}_{\mathbb{R},k_{\mathbb{R}}}) \quad \text{with} \quad \tilde{w}_{\mathbb{R},i} : \mathbb{R} \rightarrow (\mathbb{R}^{n_{\mathbb{R},i}})^{m_{\mathbb{R},i}}, \\ \tilde{w}_{\mathbb{C}} &= \text{col}(\tilde{w}_{\mathbb{C},1}, \tilde{w}_{\mathbb{C},2}, \dots, \tilde{w}_{\mathbb{C},k_{\mathbb{C}}}) \quad \text{with} \quad \tilde{w}_{\mathbb{C},i} : \mathbb{R} \rightarrow (\mathbb{C}^{n_{\mathbb{C},i}})^{m_{\mathbb{C},i}}, \\ \tilde{w}_{\mathbb{H}} &= \text{col}(\tilde{w}_{\mathbb{H},1}, \tilde{w}_{\mathbb{H},2}, \dots, \tilde{w}_{\mathbb{H},k_{\mathbb{H}}}) \quad \text{with} \quad \tilde{w}_{\mathbb{H},i} : \mathbb{R} \rightarrow (\mathbb{H}^{n_{\mathbb{H},i}})^{m_{\mathbb{H},i}}, \end{aligned}$$

as explained in the preamble.

Then $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathcal{B}) \in \mathcal{L}^q$ is ρ -symmetric if and only if there exist $m'_{\mathbb{R},i} \in \mathbb{Z}_+$, $0 \leq m'_{\mathbb{R},i} \leq m_{\mathbb{R},i}$; $m'_{\mathbb{C},i} \in \mathbb{Z}_+$, $0 \leq m'_{\mathbb{C},i} \leq m_{\mathbb{C},i}$; $m'_{\mathbb{H},i} \in \mathbb{Z}_+$, $0 \leq m'_{\mathbb{H},i} \leq m_{\mathbb{H},i}$, and polynomial matrices $A_i \in \mathbb{R}_{fr}^{m'_{\mathbb{R},i} \times m_{\mathbb{R},i}}[s]$, $C_i \in \mathbb{C}_{fr}^{m'_{\mathbb{C},i} \times m_{\mathbb{C},i}}[s]$, $H_i \in \mathbb{H}_{fr}^{m'_{\mathbb{H},i} \times m_{\mathbb{H},i}}[s]$ such that Σ admits a minimal representation

$$(11) \quad \begin{aligned} \left(A_i \left(\frac{d}{dt} \right) \otimes I_{n_{\mathbb{R},i}} \right) \tilde{w}_{\mathbb{R},i} &= 0, & i = 1, 2, \dots, k_{\mathbb{R}}, \\ \left(C_i \left(\frac{d}{dt} \right) \otimes I_{n_{\mathbb{C},i}} \right) \tilde{w}_{\mathbb{C},i} &= 0, & i = 1, 2, \dots, k_{\mathbb{C}}, \\ \left(H_i \left(\frac{d}{dt} \right) \otimes I_{n_{\mathbb{H},i}} \right) \tilde{w}_{\mathbb{H},i} &= 0, & i = 1, 2, \dots, k_{\mathbb{H}}. \end{aligned}$$

7. Applications.

7.1. Permutation symmetries.

Example 4 (simple permutations). Our first class of examples all involve the symmetric group S_q defined in Example 3, and we use the notation introduced there.

Let us now apply Theorem 5 to a system $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathcal{B}) \in \mathcal{L}^q$ with $\rho : S_q \rightarrow GL(\mathbb{R}^q)$ where $\rho_g \text{col}(w_1, w_2, \dots, w_q) := \text{col}(w_{g(1)}, w_{g(2)}, \dots, w_{g(q)})$. We obtain a suitable ρ -adapted basis by taking as new coordinates $\tilde{x}_1 := x_{av}$, $\tilde{x}_2 := \Delta x_2, \dots, \tilde{x}_q := \Delta x_q$, with $x_{av} := (1/q)(x_1 + x_2 + \dots + x_q)$ and $\Delta x_i = x_i - x_{av}$ ($i = 1, 2, \dots, q$). Clearly, $V_1 = \{\text{col}(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_q) \in \mathbb{R}^q \mid \tilde{x}_2 = \dots = \tilde{x}_q = 0\}$ and $V_2 = \{\text{col}(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_q) \in \mathbb{R}^q \mid \tilde{x}_1 = 0\}$. Observe that both the representations ρ_1 and ρ_2 introduced in Example 3 are of real type. Hence the decomposition $\rho \cong \rho_1 \oplus \rho_2$ applies to the real case. It follows that, in terms of the notation of Theorem 5, $m_{\mathbb{R},1} = m_{\mathbb{R},2} = 1$ and that all the other multiplicities are zero. Furthermore, $n_{\mathbb{R},1} = 1$ and $n_{\mathbb{R},2} = q - 1$. Thus we must consider the following choices of $m'_{\mathbb{R},i}$'s:

$$\begin{aligned} m'_{\mathbb{R},1} &= m'_{\mathbb{R},2} = 0, \\ m'_{\mathbb{R},1} &= 1; \quad m'_{\mathbb{R},2} = 0, \\ m'_{\mathbb{R},1} &= 0; \quad m'_{\mathbb{R},2} = 1, \\ m'_{\mathbb{R},1} &= m'_{\mathbb{R},2} = 1. \end{aligned}$$

The first case corresponds to the trivial situation $\mathcal{B} = C^\infty(\mathbb{R}, \mathbb{R}^q)$. In the second case, ρ -symmetry corresponds to a minimal representation of the form

$$r_{av} \left(\frac{d}{dt} \right) w_{av} = 0$$

with $w_{av} := (1/q)(w_1 + w_2 + \dots + w_q)$. This representation is determined by the nonzero polynomial $r_{av} \in \mathbb{R}[s]$. The third case yields

$$r_\Delta \left(\frac{d}{dt} \right) \Delta w_i = 0, \quad i = 2, \dots, q$$

with $\Delta w_i := w_i - w_{av}$. This representation is completely determined by the nonzero polynomial $r_\Delta \in \mathbb{R}[s]$. Note that these equations imply the redundant equation $r_\Delta (d/dt) \Delta w_1 = 0$, and hence, by letting the above equation range over $i = 1, 2, \dots, q$, we obtain an equivalent but not minimal representation. In the fourth case, we will obtain one equation on w_{av} and one on each of the Δw_i 's, and these equations are all identical.

It is clear that, by allowing nonminimal representations, all four cases can be captured in one. It follows that $\Sigma \in \mathcal{L}^q$ will be ρ -symmetric in the case of these simple permutations if and only if there exist (not necessarily nonzero) polynomials $r_{av} \in \mathbb{R}[s]$, $r_\Delta \in \mathbb{R}[s]$ such that Σ is described by

$$\begin{aligned} r_{av} \left(\frac{d}{dt} \right) w_{av} &= 0, \\ r_\Delta \left(\frac{d}{dt} \right) \Delta w_i &= 0, \quad i = 1, 2, \dots, q \end{aligned}$$

with $w_{av} := (1/q)(w_1 + w_2 + \dots + w_q)$ and $\Delta w_i := w_i - w_{av}$. Hence a symmetric system is governed by two equations. One equation governs the dynamics of the average (consider it an equation governing the *center of mass* in the case of motion of identical particles on the line). The second equation is identical for each of the components and governs the dynamics of the distance from the average and is identical for each of the components (consider this equation as governing the motion of the *displacement* of the particle from the center of mass). Note that either one or both of these equations

may be absent (when $r_{av} = 0$ and/or $r_{\Delta} = 0$). The most important feature of the above equations is the fact that the different variables w_i interact only through their average value. For an analogue nonlinear situation of this example, see [17].

Example 5 (permutations of identical subsystems with feature vectors). Next, consider the system $\Sigma = (\mathbb{R}, (\mathbb{R}^m)^n, \mathcal{B}) \in \mathcal{L}^{mn}$. Think of Σ as modelling n identical subsystems, each of which is described by m features. Thus $w = \text{col}(w_1, w_2, \dots, w_n)$ with each of the $w_i : \mathbb{R} \rightarrow \mathbb{R}^m, i = 1, 2, \dots, n$, where $w_i(t) \in \mathbb{R}^m$ denotes the feature vector of the i th subsystem at time t . In the case of the motion of n particles, this feature vector could be the position of the particle in the plane ($m = 2$) or in 3-space ($m = 3$), or we could consider each particle being described by a position and an external force acting on it (thus $m = 2, 4$, or 6 , depending on whether these particles are considered on the line, in the plane, or in 3-space).

Let $\rho : S_n \rightarrow GL((\mathbb{R}^m)^n)$ act as follows:

$$\rho_g \text{col}(w_1, w_2, \dots, w_n) := \text{col}(w_{g(1)}, w_{g(2)}, \dots, w_{g(n)}).$$

In this case, the decomposition of ρ into irreducible components leads to $\rho \cong m\rho_1 \oplus m\rho_2$ with ρ_1 and ρ_2 as in Example 3 or Example 4. The ρ -adapted basis may now be chosen as follows. Define, for $x = \text{col}(x_1, x_2, \dots, x_n) \in (\mathbb{R}^m)^n$, $x_{av} := (1/n)(x_1 + x_2 + \dots + x_n)$ and $\Delta x_i := x_i - x_{av}$. Represent x by the coordinate vector $\text{col}(x_{av}, \Delta x_2, \dots, \Delta x_n)$ and define V_1 by $\Delta x_2 = \dots = \Delta x_n = 0$ and V_2 by $x_{av} = 0$. Then $m\rho_1 \cong \rho|_{V_1}$ and $m\rho_2 \cong \rho|_{V_2}$. The further decomposition of $m\rho_1$ and $m\rho_2$ into their irreducible components is rather obvious but will not be given, since it will not be needed in the following.

In terms of the notation of Theorem 5, we have $m_{\mathbb{R},1} = m_{\mathbb{R},2} = m$ and $n_{\mathbb{R},1} = 1, n_{\mathbb{R},2} = n - 1$. Thus we should consider all the cases where $0 \leq m'_{\mathbb{R},2}, m''_{\mathbb{R},2} \leq m$. Proceeding exactly as in Example 4 we obtain that a system $\Sigma = (\mathbb{R}, (\mathbb{R}^m)^n, \mathcal{B}) \in \mathcal{L}^{mn}$ will be ρ -symmetric if and only if there exist $m_{av}, m_{\Delta} \in \mathbb{Z}_+$ (we could, but need not, restrict m_{av} and m_{Δ} to be $\leq m$) and polynomial matrices $R_{av} \in \mathbb{R}^{m_{av} \times m}[s]$ and $R_{\Delta} \in \mathbb{R}^{m_{\Delta} \times m}[s]$ such that Σ is described by

$$\begin{aligned} R_{av} \left(\frac{d}{dt} \right) w_{av} &= 0, \\ R_{\Delta} \left(\frac{d}{dt} \right) \Delta w_i &= 0, \quad i = 1, 2, \dots, n \end{aligned}$$

with $w_{av} := (1/n)(w_1 + w_2 + \dots + w_n)$ and $\Delta w_i := w_i - w_{av}$.

As a more specific example, consider a system of n identical particles in 3-space with external forces. Such a system will hence be described by differential equations of the form

$$\begin{aligned} P_{av} \left(\frac{d}{dt} \right) q_{av} &= Q_{av} \left(\frac{d}{dt} \right) F_{av}, \\ P_{\Delta} \left(\frac{d}{dt} \right) (q_i - q_{av}) &= Q_{\Delta} \left(\frac{d}{dt} \right) (F_i - F_{av}), \quad i = 1, 2, \dots, n \end{aligned}$$

with q_i the position of the i th particle, F_i the external force acting on it, and q_{av}, F_{av} defined in the obvious way. Thus the motion of the center of mass is governed by a law involving the mean force, while the laws governing the motion of the displacement from the center of mass involves the difference of the force acting on the particle and

the mean force and is identical for each of the particles. In particular, the particles interact only through the center of mass and the mean force.

Example 6 (permutations with two kinds of subsystems). Now consider the system

$$\Sigma = (\mathbb{R}, (\mathbb{R}^{m_1})^{n_1} \times (\mathbb{R}^{m_2})^{n_2}, \mathcal{B}) \in \mathcal{L}^{m_1 n_1 + m_2 n_2}.$$

Think of Σ as modelling n_1 identical subsystems of one kind with each m_1 features and n_2 identical subsystems of a second kind each with m_2 features.

Let $\rho : S_{n_1} \times S_{n_2} \rightarrow GL((\mathbb{R}^{m_1})^{n_1} \times (\mathbb{R}^{m_2})^{n_2})$ act as follows:

$$\rho_{g_1, g_2} \text{col}(w'_1, \dots, w'_{n_1}, w''_1, \dots, w''_{n_2}) := \text{col}(w'_{g_1(1)}, \dots, w'_{g_1(n_1)}, w''_{g_2(1)}, \dots, w''_{g_2(n_2)}).$$

The decomposition of ρ into irreducible components now becomes $\rho \cong (m_1 + m_2)\rho_1 \oplus m_1\rho'_2 \oplus m_2\rho''_2$ with ρ_1 the identity representation and $(m_1 + m_2)\rho_1 \cong \rho|_{V_1}$ with

$$V_1 = \{\text{col}(x'_1, \dots, x'_{n_1}, x''_1, \dots, x''_{n_2}) \mid x'_1 = \dots = x'_{n_1}, x''_1 = \dots = x''_{n_2}\}.$$

Furthermore, ρ'_2 and ρ''_2 are the analogues of what we denoted by ρ_2 before: ρ'_2 corresponds to the analogue of ρ_2 as an irreducible representation of S_{n_1} and ρ''_2 corresponds to the analogue of ρ_2 as an irreducible representation of S_{n_2} . Proceeding as before, Theorem 5 will show that Σ is ρ -symmetric if and only if there exist $m_{av}, m'_\Delta, m''_\Delta \in \mathbb{Z}_+$ and polynomial matrices $R_{av} \in \mathbb{R}^{m_{av} \times (m_1 + m_2)}[s]$, $R'_\Delta \in \mathbb{R}^{m'_\Delta \times m_1}[s]$, and $R''_\Delta \in \mathbb{R}^{m''_\Delta \times m_2}[s]$ such that Σ is described by

$$\begin{aligned} R_{av} \left(\frac{d}{dt} \right) \begin{bmatrix} w'_{av} \\ \dots \\ w''_{av} \end{bmatrix} &= 0, \\ R'_\Delta \left(\frac{d}{dt} \right) (w'_i - w'_{av}) &= 0, \quad i = 1, 2, \dots, n_1, \\ R''_\Delta \left(\frac{d}{dt} \right) (w''_j - w''_{av}) &= 0, \quad j = 1, 2, \dots, n_2. \end{aligned}$$

This shows that the two groups can only interact through their averages, while all the respective displacements are independent.

An interesting special case can now be obtained by taking $n_2 = 1$. We can then view the situation as modelling the dynamics of the interaction of a central control station with n identical substations. In the obvious notation, the dynamical laws then take the form

$$\begin{aligned} R_{av} \left(\frac{d}{dt} \right) \begin{bmatrix} w_{av} \\ \dots \\ w_{\text{central}} \end{bmatrix} &= 0, \\ R_\Delta \left(\frac{d}{dt} \right) (w_i - w_{av}) &= 0, \quad i = 1, 2, \dots, n. \end{aligned}$$

Thus the central controller influences only the average of the feature vectors of the substations.

Remark. The above examples all involve the action of the whole symmetric group S_n . Actually, identical results may be obtained by considering a doubly transitive subgroup G of S_n . A subgroup $G \subseteq S_n$ is said to be *transitive* if, for all $\alpha, \beta \in \{1, 2, \dots, n\}$, there exist $g \in G$ such that $g(\alpha) = \beta$; it is said to be *doubly transitive*

if, for all $\alpha', \beta', \alpha'', \beta'' \in \{1, 2, \dots, n\}$, $\alpha' \neq \alpha''$, and $\beta' \neq \beta''$, there exists $g \in G$ such that $g(\alpha') = \beta'$ and $g(\alpha'') = \beta''$. An example of a doubly transitive subgroup is the subgroup of even permutations.

It can be shown that the representations of Examples 4–6 remain valid without changes if we assume invariance of \mathcal{B} for a doubly transitive subgroup G of S_n in Example 5 and doubly transitive subgroups G_1 of S_{n_1} and G_2 of S_{n_2} in Example 6. In particular, this shows that permutation symmetry for one doubly transitive subgroup implies permutation symmetry for the whole of S_n in Example 5 and analogously for the whole of $S_{n_1} \times S_{n_2}$ in Example 6!

Example 7 (cyclic permutations). Let \mathbb{Z}_q denote the group consisting of

$$\{0, 1, \dots, q-1\}$$

with the group operation addition modulo q . It is more convenient to denote this group as $\{1, r, r^2, \dots, r^{q-1}\}$ with $r^q = 1$. Note that \mathbb{Z}_q is a subgroup of S_q . It is called the group of cyclic permutations. Let $\mathbb{Z}_q = \{1, r, \dots, r^{q-1}\}$ act on \mathbb{K}^q as in the case of permutations. Thus

$$\rho_r \text{col}(x_1, x_2, \dots, x_{q-1}, x_q) := \text{col}(x_q, x_1, x_2, \dots, x_{q-1})$$

from which $\rho_r, \dots, \rho_{r^{q-1}}$ follow.

Because the group \mathbb{Z}_q is commutative, its irreducible representations over \mathbb{C} are all one-dimensional. There are q such irreducible representations given by $\rho_1, \rho_2, \dots, \rho_q$ with $\rho_k : \mathbb{Z}_q \rightarrow GL(\mathbb{C})$ given by $\rho_{k,r} = \lambda^k$ with $\lambda := e^{i(2\pi/q)}$. A simple calculation shows that $\rho \cong \rho_1 \oplus \rho_2 \oplus \dots \oplus \rho_q$, with the invariant subspace corresponding to ρ_k given by $\text{span col}(1, \lambda^{-k}, \lambda^{-2k}, \dots, \lambda^{-(q-1)k})$. This now allows us to compute the ρ -adapted basis. We omit the detailed calculations.

Theorem 4 implies that $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathcal{B}) \in \mathcal{L}^q$ will be symmetric with respect to the cyclic permutations if and only if there exist polynomials $r_1, r_2, \dots, r_q \in \mathbb{C}[z]$, $p(\Sigma)$ of which are nonzero, such that

$$r_j \left(\frac{d}{dt} \right) \left(\sum_{k=1}^q \lambda^{-jk} w_k \right) = 0, \quad j = 1, 2, \dots, q$$

forms a minimal representation of Σ . This set of equations can be edited a bit further and leads to the following canonical form for this cyclic symmetry. Σ will be symmetric if and only if there exist polynomials $\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_q \in \mathbb{C}[s]$ and a representation of Σ of the following form:

$$\begin{bmatrix} \tilde{r}_1 \left(\frac{d}{dt} \right) & \tilde{r}_2 \left(\frac{d}{dt} \right) & \cdots & \tilde{r}_{q-1} \left(\frac{d}{dt} \right) & \tilde{r}_q \left(\frac{d}{dt} \right) \\ \tilde{r}_q \left(\frac{d}{dt} \right) & \tilde{r}_1 \left(\frac{d}{dt} \right) & \cdots & \tilde{r}_{q-2} \left(\frac{d}{dt} \right) & \tilde{r}_{q-1} \left(\frac{d}{dt} \right) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{r}_2 \left(\frac{d}{dt} \right) & \tilde{r}_3 \left(\frac{d}{dt} \right) & \cdots & \tilde{r}_q \left(\frac{d}{dt} \right) & \tilde{r}_1 \left(\frac{d}{dt} \right) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_q \end{bmatrix} = 0.$$

Note that this representation (which is possibly nonminimal) puts the cyclic symmetry nicely into evidence.

In the real case where $\mathbb{K} = \mathbb{R}$, we must combine the complex conjugate irreducible representations. Similar calculations to the above ones lead to (possibly nonminimal)

behavioral equations

$$\begin{aligned}
 r_0 \left(\frac{d}{dt} \right) \left(\sum_{k=1}^q w_k \right) &= 0, \\
 r'_j \left(\frac{d}{dt} \right) \left(\sum_{k=1}^q w_k \cos \frac{2\pi j}{q} k \right) + r''_j \left(\frac{d}{dt} \right) \left(\sum_{k=1}^q w_k \sin \frac{2\pi j}{q} k \right) &= 0, \\
 r''_j \left(\frac{d}{dt} \right) \left(\sum_{k=1}^q w_k \cos \frac{2\pi j}{q} k \right) - r'_j \left(\frac{d}{dt} \right) \left(\sum_{k=1}^q w_k \sin \frac{2\pi j}{q} k \right) &= 0, \\
 &\text{for } j = 1, 2, \dots, \frac{q-1}{2}, \quad \text{if } q \text{ is odd,} \\
 &\text{for } j = 1, 2, \dots, \frac{q}{2} - 1, \quad \text{if } q \text{ is even.}
 \end{aligned}$$

In the case that q is even, there is an additional equation

$$r_{\frac{q}{2}} \left(\frac{d}{dt} \right) \left(\sum_{k=1}^q (-1)^k w_k \right) = 0,$$

where $r_0, r'_j, r''_j, r_{q/2} \in \mathbb{R}[s]$.

The above calculations are easily generalized to cyclic permutation symmetries with feature vectors. In this case, it suffices to interpret the r 's as matrices.

7.2. A quaternionic symmetry.

Example 8. Consider $(\mathbb{R}, \mathbb{R}^q, \mathcal{B}) \in \mathcal{L}^q$ with $q = 4$. Now assume that this system is symmetric in the following sense:

$$\left(\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \in \mathcal{B} \right) \Rightarrow \left(\begin{bmatrix} -w_2 \\ w_1 \\ -w_4 \\ w_3 \end{bmatrix}, \begin{bmatrix} -w_3 \\ w_4 \\ w_1 \\ -w_2 \end{bmatrix}, \begin{bmatrix} -w_4 \\ -w_3 \\ w_2 \\ w_1 \end{bmatrix} \in \mathcal{B} \right).$$

As we will see, this is a quaternionic symmetry. We will not give a physical example where such a symmetry can occur.

The group of quaternions consists of $\mathcal{H} = \{\pm 1, \pm i, \pm j, \pm k\}$ with multiplication table

$$i^2 = j^2 = k^2 = -1, \quad ij = -ji = k, \quad jk = -kj = i, \quad ki = -ik = -j.$$

The following defines a representation of \mathcal{H} on \mathbb{C}^2

$$\begin{aligned}
 \pm 1 &\mapsto \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & \pm i &\mapsto \pm \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \\
 \pm j &\mapsto \pm \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, & \pm k &\mapsto \pm \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.
 \end{aligned}$$

This representation is irreducible. It is obviously not real and, since all the above matrices have real trace, it is a complex representation of quaternionic type. The

representation induced on \mathbb{R}^4 by combining this representation with its complex conjugate yields

$$\begin{aligned} 1 &\mapsto \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, & i &\mapsto M_1 := \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \\ j &\mapsto M_2 := \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, & k &\mapsto M_3 := \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Note that M_1, M_2, M_3 are precisely the matrices defining the static symmetry under consideration.

It follows from Theorem 5 that the Σ 's that are symmetric in this sense are precisely those that admit a representation for the form

$$\begin{aligned} a \left(\frac{d}{dt} \right) w_1 - b \left(\frac{d}{dt} \right) w_2 - c \left(\frac{d}{dt} \right) w_3 - d \left(\frac{d}{dt} \right) w_4 &= 0, \\ b \left(\frac{d}{dt} \right) w_1 + a \left(\frac{d}{dt} \right) w_2 + d \left(\frac{d}{dt} \right) w_3 - c \left(\frac{d}{dt} \right) w_4 &= 0, \\ c \left(\frac{d}{dt} \right) w_1 - d \left(\frac{d}{dt} \right) w_2 + a \left(\frac{d}{dt} \right) w_3 + b \left(\frac{d}{dt} \right) w_4 &= 0, \\ d \left(\frac{d}{dt} \right) w_1 + c \left(\frac{d}{dt} \right) w_2 - b \left(\frac{d}{dt} \right) w_3 + a \left(\frac{d}{dt} \right) w_4 &= 0 \end{aligned}$$

with $a, b, c, d \in \mathbb{R}[s]$.

7.3. Symmetries with Lie groups. A Lie group G is a topological group with the structure of a C^∞ differentiable \mathbb{K} -manifold in which the group multiplication and the inverse are C^∞ maps. A C^∞ group homomorphism $\rho : G \rightarrow GL(\mathbb{K}^n)$ is said to be a representation of the Lie group G on \mathbb{K}^n . Let $gl(\mathbb{K}^n)$ denote the set of $n \times n$ matrices over \mathbb{K} endowed with the usual commutator product $[A, B] = AB - BA$; $gl(\mathbb{K}^n)$ is the Lie algebra of $GL(\mathbb{K}^n)$. Let \mathcal{G} be the Lie algebra of G . The representation $\rho : G \rightarrow GL(\mathbb{K}^n)$ of the Lie group G on \mathbb{K}^n induces a Lie algebra homomorphism $\tilde{\rho} : \mathcal{G} \rightarrow gl(\mathbb{K}^n)$ of the Lie algebra \mathcal{G} on \mathbb{K}^n . Let $X_1, X_2, \dots, X_N \in \mathbb{K}^{n \times n}$ be a set of generators of $\tilde{\rho}(\mathcal{G})$. Then we have the following result.

THEOREM 6. *Let \mathcal{G} be a compact connected Lie group and let $\rho : G \rightarrow GL(\mathbb{K}^q)$ be a representation of the Lie group \mathcal{G} on \mathbb{K}^q . Let $X_1, X_2, \dots, X_N \in \mathbb{K}^{q \times q}$ be a set of generators of $\tilde{\rho}(\mathcal{G})$ with $\tilde{\rho}$ the induced representation of the Lie algebra \mathcal{G} on \mathbb{K}^q . Then the following are equivalent for $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$:*

- (1) Σ is ρ -symmetric;
- (2) $XB \subseteq \mathcal{B}$ for all $X \in \tilde{\rho}(\mathcal{G})$;
- (3) $X_i \mathcal{B} \subseteq \mathcal{B}$ for $i = 1, 2, \dots, N$;
- (4) *There exist matrices $Y_1, Y_2, \dots, Y_n \in \mathbb{K}^{p(\Sigma) \times p(\Sigma)}$ and a $R \in \mathbb{K}^{p(\Sigma) \times q[s]}$ such that $R(d/dt)w = 0$ is a minimal representation for Σ with $Y_i R(s) = R(s)X_i$ for $i = 1, 2, \dots, N$. Moreover, the subspace generated by the Y_i 's is a Lie subalgebra of $gl(\mathbb{K}^{p(\Sigma)})$ and yields through the association $X_i \mapsto Y_i$ a subrepresentation of the Lie algebra representation $\tilde{\rho} : \mathcal{G} \rightarrow gl(\mathbb{K}^n)$.*

Proof. Assume without loss of generality that G is a subgroup of $GL(\mathbb{K}^q)$ and that ρ is the canonical injection.

We will run the circle $(4) \Rightarrow (3) \Rightarrow (2) \Rightarrow (1) \Rightarrow (4)$. The first two implications are trivial.

To show that $(2) \Rightarrow (1)$, assume that $X \in \mathcal{G}$ and consider the exponential $\exp X$. It is well known that there exist $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{K}$ such that $\exp X = \sum_{i=1}^n \alpha_i X^i$. This implies, by (2), that $(\exp X)\mathcal{B} \subseteq \mathcal{B}$. Since G is compact connected, $\exp : \mathcal{G} \rightarrow G$ is surjective, and hence (1) follows.

$(1) \Rightarrow (4)$: From Theorem 3, it follows that there exist $R \in \mathbb{K}^{p(\Sigma) \times q}[s]$ and a representation $\rho' : G \rightarrow GL(\mathbb{K}^{p(\Sigma)})$ such that $\rho' R(s) = R(s)\rho$. Now consider the one-parameter subgroup of G given by $G_i := \{\exp \mu X_i \mid \mu \in \mathbb{K}\}$. Then $\rho'(G_i)$ is also a one-parameter subgroup of $GL(\mathbb{K}^{p(\Sigma)})$. Hence there exists a $Y_i \in \mathfrak{gl}(\mathbb{K}^{p(\Sigma)})$ such that $\rho'(\exp \mu X_i) = \exp \mu Y_i$. Hence $(\exp \mu Y_i)R(s) = R(s)(\exp \mu X_i)$. Differentiating at $\mu = 0$ yields $Y_i R(s) = R(s)X_i$. The last part of (4) follows from the observation that $R(s)X_i X_j = Y_i R(s)X_j = Y_i Y_j R(s)$. Hence $R(s)[X_i, X_j] = [Y_i, Y_j]R(s)$. This shows that the vector space generated by these Y_i 's is a Lie subalgebra of $\mathfrak{gl}(\mathbb{K}^{p(\Sigma)})$ and that $X_i \mapsto Y_i$ generates a Lie algebra homomorphism. \square

7.4. Rotation symmetries.

Example 9 (rotations on \mathbb{R}^m with $m > 2$). Consider the group $SO(m)$ of real orthogonal $m \times m$ matrices with determinant 1. $SO(m)$ is a compact group, and it can be shown that, when $m \geq 3$, the canonical injection of $SO(m)$ into $GL(\mathbb{C}^m)$ (which will also be denoted as $SO(m)$) is irreducible over \mathbb{C} .

Now consider the system $\Sigma = (\mathbb{R}, (\mathbb{R}^m)^n, \mathcal{B})$. Assume that $SO(m)$ acts on $(\mathbb{R}^m)^n$ by $M \in SO(m)$, taking $\text{col}(w_1, w_2, \dots, w_n)$ into $\text{col}(Mw_1, Mw_2, \dots, Mw_n)$. Now consider the static symmetry induced by this action. To interpret this situation physically, think, for example, of $m = 3$ and of Σ as modelling the motion of n (not identical) particles in \mathbb{R}^3 under the influence of rotation invariant laws. Another possibility is to think of the situation $m = 3$ and $n = 2n'$ with Σ modelling the position and the force acting on n' particles.

Applying Theorem 5 and using the irreducibility of $SO(m)$ for $m > 2$ immediately shows that Σ will be symmetric if and only if there exists a polynomial matrix $R' \in \mathbb{R}^{n \times n}[s]$ such that Σ is represented as

$$\left(R' \left(\frac{d}{dt} \right) \otimes I_m \right) \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = 0.$$

Example 10 (rotations on \mathbb{R}^2). The above example must be modified in the case where $m = 2$, since the canonical injection of $SO(2)$ into $GL(\mathbb{C}^2)$ is reducible over \mathbb{C} . $SO(2)$ can be written as $SO(2) \cong \rho \oplus \rho^*$ with

$$\rho \left(\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \right) = e^{i\theta}.$$

Applying Theorem 5 leads to the following representation of rotation symmetric systems for $m = 2$:

$$\begin{bmatrix} R_1 \left(\frac{d}{dt} \right) & -R_2 \left(\frac{d}{dt} \right) \\ R_2 \left(\frac{d}{dt} \right) & R_1 \left(\frac{d}{dt} \right) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0$$

with $R_1, R_2 \in \mathbb{R}^{(1/2)p(\Sigma) \times n}$.

This result can also be obtained from Theorem 6. Indeed, since $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ generates $SO(2)$ through exponentiation, symmetry implies the existence of an R such that

$$\begin{bmatrix} 0_{(1/2)p(\Sigma)} & I_{(1/2)p(\Sigma)} \\ -I_{(1/2)p(\Sigma)} & 0_{(1/2)p(\Sigma)} \end{bmatrix} R(s) = R(s) \begin{bmatrix} 0_n & I_n \\ -I_n & 0_n \end{bmatrix},$$

where I_k and O_k denote the $k \times k$ identity and zero matrices, respectively. This also yields the above representation.

Finally, systems with this rotation symmetry can also be represented by introducing the complex signal $w := w_1 + iw_2$ with $w : \mathbb{R} \rightarrow \mathbb{C}$. The differential equation governing

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

then becomes $R(d/dt)w = 0$ with $R \in \mathbb{C}^{(1/2)p(\Sigma) \times n}[s]$.

Example 11 (rotations over $2\pi/n$ degrees on \mathbb{R}^2). Let $G = \mathbb{Z}_n = \{0, 1, \dots, n-1\}$ (see Example 7 for notation) with $n \geq 3$. Now consider the representation ρ of G on \mathbb{R}^2 , defined by

$$\rho : k \mapsto \begin{bmatrix} \cos \frac{2\pi k}{n} & -\sin \frac{2\pi k}{n} \\ \sin \frac{2\pi k}{n} & \cos \frac{2\pi k}{n} \end{bmatrix}.$$

Then $(\mathbb{R}, \mathbb{R}^2, \mathcal{B}) \in \mathcal{L}^2$ being ρ -symmetric means symmetry in the sense of Example 10, by rotations of $2\pi/n$ degrees. This is a subgroup of $SO(2)$. Since the representation ρ is also irreducible, we immediately obtain from Theorem 5 the representation obtained in Example 10. This allows the interesting conclusion that $\Sigma \in \mathcal{L}^2$ being symmetric with respect to rotations of $2\pi/n$ degrees will imply that it is symmetric with respect to *all* rotations. The Lie algebra line of reasoning in Example 10 showed this already for $n = 4$, that is, for 90° rotations.

8. State space models. Many useful models encountered in applications involve auxiliary variables in addition to the variables that the model aims at describing. To distinguish between these two kinds of variables, we call the variables of primary interest *manifest* and denote them as w , and we call the auxiliary variables *latent*, and—usually—denote them as ℓ .

Proceeding with the terminology of [15], this leads to a *dynamical system with latent variables*, defined as $\Sigma_f = (\mathbb{T}, \mathbb{W}, \mathbb{L}, \mathcal{B}_f)$ with $\mathbb{T} \subseteq \mathbb{R}$ the *time axis*, \mathbb{W} the signal space of *manifest variables*, \mathbb{L} the signal space of *latent variables*, and $\mathcal{B}_f \subseteq (\mathbb{W} \times \mathbb{L})^{\mathbb{T}}$ the *full behavior*. Σ_f induces the *manifest* dynamical system $\Sigma = (\mathbb{T}, \mathbb{W}, \mathcal{B})$ with *manifest behavior* $\mathcal{B} = \{w \mid \text{there exists an } \ell \text{ such that } (w, \ell) \in \mathcal{B}_f\}$.

In the context of systems described by differential equations, this leads us to consider linear differential equations

$$(12) \quad R\left(\frac{d}{dt}\right)w = M\left(\frac{d}{dt}\right)\ell$$

linking the manifest variables $w \in C^\infty(\mathbb{R}, \mathbb{K}^q)$ to the latent variables $\ell \in C^\infty(\mathbb{R}, \mathbb{K}^d)$. Here $R \in \mathbb{K}^{\bullet \times q}[s]$ and $M \in \mathbb{K}^{\bullet \times d}[s]$ are two polynomial matrices with the same number of rows. Formally, (12) defines the latent variable dynamical system $(\mathbb{R}, \mathbb{K}^q, \mathbb{K}^d, \mathcal{B}_f)$

with $\mathcal{B}_f = \ker[R(d/dt) \vdash M(d/dt)]$. Obviously, $(\mathbb{R}, \mathbb{K}^{q+d}, \mathcal{B}_f) \in \mathcal{L}^{q+d}$. We will denote this class of latent variable dynamical systems as $\mathcal{L}^{q,d}$. This family of models is hence parametrized by pairs of polynomial matrices (R, M) .

This model with latent variables induces the manifest dynamical system $(\mathbb{R}, \mathbb{K}^q, \mathcal{B})$ with $\mathcal{B} = (R(d/dt))^{-1} \text{im } M(d/dt)$. It can be shown that $(\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$, that is, that $(R(d/dt))^{-1} \text{im } M(d/dt)$ can itself be described as the solution set of a system of constant coefficient differential equations. In other words, there exists a polynomial matrix $R' \in \mathbb{R}^{\bullet \times q}[s]$ such that the manifest behavior of (12) is described by

$$(13) \quad R' \left(\frac{d}{dt} \right) w = 0.$$

The matrix R' can be obtained as follows. By premultiplying M by a suitable unimodular polynomial matrix U , UM can be brought in the form

$$UM = \begin{bmatrix} 0 \\ \dot{M}'' \end{bmatrix}$$

with $M'' \in \mathbb{K}^{\bullet \times d}_{fr}[s]$. Partitioning UR conformably as

$$UR = \begin{bmatrix} R' \\ \dot{R}'' \end{bmatrix}$$

yields the desired R' . The result that (13) then defines the manifest behavior of (12) follows easily from the observation that $M''(d/dt)$ is surjective (in other words, if $P \in \mathbb{R}^{n_1 \times n_2}_{fr}[s]$, then $P(d/dt) : C^\infty(\mathbb{R}; \mathbb{K}^{n_2}) \rightarrow C^\infty(\mathbb{R}; \mathbb{K}^{n_1})$ will be a surjective map).

A very useful class of systems with latent variables are the *state space systems*. In this case, the latent variables are denoted by x instead of by ℓ . In [15], [16], we have defined state space models abstractly in terms of concatenation of trajectories. For differential systems, the result is that state space systems are precisely those latent variable systems whose full behavior can be described by the following special type of differential equations linking the state trajectory $x \in C^\infty(\mathbb{R}; \mathbb{K}^n)$ to the manifest trajectory $w \in C^\infty(\mathbb{R}; \mathbb{K}^q)$:

$$(14) \quad \left(E \frac{d}{dt} + F \right) x + Hw = 0$$

with E, F, H matrices over \mathbb{K} of suitable dimension. The crucial feature of (14) is that this differential equation is first-order in x and zeroth-order in w .

Let us denote the state space systems with manifest signal space \mathbb{K}^q , and state space \mathbb{K}^n by $\mathcal{L}^{q,n}_s$. It follows that $\mathcal{L}^{q,n}_s$ is parametrized by $\mathbb{K}^{\bullet \times (2n+q)}$ by associating with the element $[E:F:H] \in \mathbb{K}^{\bullet \times (2n+q)}$ the behavioral equations (14). We denote the state space system described by (14) simply as (E, F, H) .

It follows immediately from the elimination of latent variables that the manifest behavior of $\Sigma_s \in \mathcal{L}^{q,n}_s$ leads to a system in \mathcal{L}^q . However, the converse also holds: for any $\Sigma \in \mathcal{L}^q$, i.e., for any polynomial matrix $R \in \mathbb{K}^{\bullet \times q}[s]$, there exist nonnegative integers $n, f \in \mathbb{N}_0$ and matrices $E, F \in \mathbb{K}^{f \times n}$, $H \in \mathbb{K}^{f \times q}$ such that the manifest behavior of (14) is represented by (1). If $\Sigma_s \in \mathcal{L}^{q,n}_s$ induces in this sense the system $\Sigma \in \mathcal{L}^q$, then we call Σ_s (or $(E, F, H) \in \mathbb{K}^{\bullet \times (2n+q)}$) a *state space representation* of $\Sigma \in \mathcal{L}^q$ (or of (1)). We denote this as $(E, F, H) \rightsquigarrow \Sigma$, or $(E, F, H) \rightsquigarrow R$.

Let $E, F \in \mathbb{K}^{f \times n}$, $H \in \mathbb{K}^{f \times q}$, and $(E, F, H) \rightsquigarrow \Sigma \in \mathcal{L}^q$. We call this state space system *minimal* if $(E', F' \in \mathbb{K}^{f' \times n}, H' \in \mathbb{K}^{f' \times q}$, and $(E', F', H') \rightsquigarrow \Sigma$) implies $(f \leq f'$ and $n \leq n')$. In [15], [16], it is shown that any system $\Sigma \in \mathcal{L}^q$ admits a minimal state space representation (in other words, both f and n can be simultaneously minimized). This implies that, in a minimal state representation, the number of state variables will depend on Σ , but not on the particular minimal state space representation of Σ . We denote this minimal number of state variables by $n(\Sigma)$. Similarly, $f(\Sigma)$ denotes the minimal number of equations in (14) representing Σ . It is easy to see [15] that $f(\Sigma) = n(\Sigma) + p(\Sigma)$. The following proposition play a crucial role in our proof of Theorem 3.

PROPOSITION 7. *Let $(E, F, H) \rightsquigarrow \Sigma \in \mathcal{L}^q$ be minimal. Then $(E', F', H') \rightsquigarrow \Sigma$ will also be minimal if and only if there exist $V \in GL(\mathbb{K}^{f(\Sigma)})$ and $T \in GL(\mathbb{K}^{n(\Sigma)})$ such that*

$$E' = VET, \quad F' = VFT, \quad \text{and} \quad H' = VH.$$

Finally, for given (E, F, H) and (E', F', H') , this V and T are unique.

Proof. For the proof, see [15]. \square

This proposition states that minimal representations of a given manifest behavior differ only in their choice of the basis in the state space and in the equation space.

THEOREM 8. *Let $\rho : G \rightarrow GL(\mathbb{K}^q)$ be a representation of the compact group G on \mathbb{K}^q . Let $\Sigma \in \mathcal{L}^q$ and let $\Sigma_s \cong (E, F, H) \in \mathcal{L}^{q,n}$ be a minimal state space representation of Σ . Then Σ is ρ -symmetric if and only if there exist representations $\rho' : G \rightarrow GL(\mathbb{K}^{n(\Sigma)})$ and $\rho'' : G \rightarrow GL(\mathbb{K}^{f(\Sigma)})$ such that*

$$(15) \quad \rho''E = E\rho', \quad \rho''F = F\rho', \quad \text{and} \quad \rho''H = H\rho.$$

Proof. The “if” part is clear. To see the “only if” part, observe that $(E, F, H\rho_g)$ is also a minimal state space representation of Σ . By Proposition 7, this implies that there exist $V_g \in GL(\mathbb{K}^{f(\Sigma)})$ and $T_g \in GL(\mathbb{K}^{n(\Sigma)})$ such that $E = V_gET_g$, $F = V_gFT_g$, and $H\rho_g = V_gH$. Define $\rho' : g \mapsto V_g$ and $\rho'' : g \mapsto (T_g)^{-1}$. We claim that ρ' and ρ'' are also representations of G . In fact, from the uniqueness condition in Proposition 7, it follows immediately that $V_{g_1g_2} = V_{g_1}V_{g_2}$ and $(T_{g_1g_2})^{-1} = (T_{g_1})^{-1}(T_{g_2})^{-1}$. \square

9. Canonical forms for state space models of symmetric systems. We now use Theorem 8 and the ideas of §6 to obtain canonical forms for state space systems. It is convenient to distinguish again between the complex case where $\mathbb{K} = \mathbb{C}$ and the real case where $\mathbb{K} = \mathbb{R}$.

9.1. Complex state space systems. Write the representations ρ, ρ' , and ρ'' of Theorem 8 in terms of irreducible representations as

$$\begin{aligned} \rho &\cong m_1\rho_1 \oplus m_2\rho_2 \oplus \cdots \oplus m_k\rho_k, \\ \rho' &\cong m'_1\rho_1 \oplus m'_2\rho_2 \oplus \cdots \oplus m'_k\rho_k, \\ \rho'' &\cong m''_1\rho_1 \oplus m''_2\rho_2 \oplus \cdots \oplus m''_k\rho_k. \end{aligned}$$

Now choose a ρ -adapted basis in the manifest signal space \mathbb{C}^q , a ρ' -adapted basis in the state space $\mathbb{C}^{n(\Sigma)}$, and a ρ'' -adapted basis in the equation space $\mathbb{C}^{f(\Sigma)}$. Now apply Schur’s lemma to (15). This yields that, in these bases E, F , and H will take the

form

$$\begin{aligned} E &= \text{diag}(E_1 \otimes I_{n_1}, E_2 \otimes I_{n_2}, \dots, E_k \otimes I_{n_k}), \\ F &= \text{diag}(F_1 \otimes I_{n_1}, F_2 \otimes I_{n_2}, \dots, F_k \otimes I_{n_k}), \\ H &= \text{diag}(H_1 \otimes I_{n_1}, H_2 \otimes I_{n_2}, \dots, H_k \otimes I_{n_k}). \end{aligned}$$

This yields the following result.

THEOREM 9. *Let G be a compact group and let $\rho : GL(\mathbb{C}^q)$ be a representation of G on \mathbb{C}^q . Assume that $\rho \cong m_1\rho_1 \oplus m_2\rho_2 \oplus \dots \oplus m_k\rho_k$ with $\rho_i : G \rightarrow GL(\mathbb{C}^{n_i})$, $i = 1, 2, \dots, k$, distinct irreducible representations. Assume that the basis in \mathbb{C}^q is ρ -adapted, as in Theorem 4. Then $\Sigma = (\mathbb{R}, \mathbb{C}^q, \mathcal{B}) \in \mathcal{L}^q$ is ρ -symmetric if and only if there exist $m'_i, m''_i \in \mathbb{Z}_+$ and matrices $E_i, F_i \in \mathbb{C}^{m'_i \times m'_i}, H_i \in \mathbb{C}^{m''_i \times m_i}$ such that Σ admits the minimal state space representation of the form*

$$(16) \quad \left(\left(E_i \frac{d}{dt} + F_i \right) \otimes I_{n_i} \right) x_i + (H_i \otimes I_{n_i}) \tilde{w}_i = 0, \quad i = 1, 2, \dots, k$$

with \tilde{w} as in Theorem 4, and where $x_i : \mathbb{R} \rightarrow (\mathbb{C}^{n_i})^{m'_i}$, $x = \text{col}(x_1, x_2, \dots, x_k)$, is the state trajectory.

9.2. Real state space systems. To obtain a canonical form for real symmetric state space systems, we proceed in complete analogy to §6.2. By considering the components of real, complex, and quaternionic type in the decomposition of ρ , the following result is obtained.

THEOREM 10. *Let G be a compact group and let $\rho : G \rightarrow GL(\mathbb{R}^q)$ be a representation of G on \mathbb{R}^q . Assume that ρ is decomposed as in (8) and that the basis in \mathbb{R}^q is ρ -adapted, as in Theorem 5. Then $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathcal{B}) \in \mathcal{L}^q$ is ρ -symmetric if and only if there exist*

$$m'_{\mathbb{R},i}, m'_{\mathbb{C},i}, m'_{\mathbb{H},i}, m''_{\mathbb{R},i}, m''_{\mathbb{C},i}, m''_{\mathbb{H},i} \in \mathbb{Z}_+$$

and matrices

$$E_{\mathbb{R},i}, F_{\mathbb{R},i} \in \mathbb{R}^{m'_{\mathbb{R},i} \times m'_{\mathbb{R},i}}, H_{\mathbb{R},i} \in \mathbb{R}^{m''_{\mathbb{R},i} \times m_{\mathbb{R},i}};$$

$$E_{\mathbb{C},i}, F_{\mathbb{C},i} \in \mathbb{C}^{m'_{\mathbb{C},i} \times m'_{\mathbb{C},i}}, H_{\mathbb{C},i} \in \mathbb{C}^{m''_{\mathbb{C},i} \times m_{\mathbb{C},i}};$$

$$E_{\mathbb{H},i}, F_{\mathbb{H},i} \in \mathbb{H}^{m'_{\mathbb{H},i} \times m'_{\mathbb{H},i}}, H_{\mathbb{H},i} \in \mathbb{C}^{m''_{\mathbb{H},i} \times m_{\mathbb{H},i}},$$

such that Σ admits a minimal state space representation of the form

$$(17) \quad \begin{aligned} &\left(\left(E_{\mathbb{R},i} \frac{d}{dt} + F_{\mathbb{R},i} \right) \otimes I_{n_{\mathbb{R},i}} \right) x_{\mathbb{R},i} + (H_{\mathbb{R},i} \otimes I_{n_{\mathbb{R},i}}) \tilde{w}_{\mathbb{R},i} = 0, \quad i = 1, 2, \dots, k_{\mathbb{R}}, \\ &\left(\left(E_{\mathbb{C},i} \frac{d}{dt} + F_{\mathbb{C},i} \right) \otimes I_{n_{\mathbb{C},i}} \right) x_{\mathbb{C},i} + (H_{\mathbb{C},i} \otimes I_{n_{\mathbb{C},i}}) \tilde{w}_{\mathbb{C},i} = 0, \quad i = 1, 2, \dots, k_{\mathbb{C}}, \\ &\left(\left(E_{\mathbb{H},i} \frac{d}{dt} + F_{\mathbb{H},i} \right) \otimes I_{n_{\mathbb{H},i}} \right) x_{\mathbb{H},i} + (H_{\mathbb{H},i} \otimes I_{n_{\mathbb{H},i}}) \tilde{w}_{\mathbb{H},i} = 0, \quad i = 1, 2, \dots, k_{\mathbb{H}} \end{aligned}$$

with \tilde{w} as in Theorem 5, and where $x = \text{col}(x_{\mathbb{R}}, x_{\mathbb{C}}, x_{\mathbb{H}})$ is the state trajectory;

$$x_{\mathbb{R}} = \text{col}(x_{\mathbb{R},1}, x_{\mathbb{R},2}, \dots, x_{\mathbb{R},k_{\mathbb{R}}}), \quad \text{with } x_{\mathbb{R},i} : \mathbb{R} \rightarrow (\mathbb{R}^{n_{\mathbb{R},i}})^{m'_{\mathbb{R},i}};$$

$$x_{\mathbb{C}} = \text{col}(x_{\mathbb{C},1}, x_{\mathbb{C},2}, \dots, x_{\mathbb{C},k_{\mathbb{C}}}), \quad \text{with } x_{\mathbb{C},i} : \mathbb{R} \rightarrow (\mathbb{C}^{n_{\mathbb{C},i}})^{m'_{\mathbb{C},i}};$$

$$x_{\mathbb{H}} = \text{col}(x_{\mathbb{H},1}, x_{\mathbb{H},2}, \dots, x_{\mathbb{H},k_{\mathbb{H}}}), \quad \text{with } x_{\mathbb{H},i} : \mathbb{R} \rightarrow (\mathbb{H}^{n_{\mathbb{H},i}})^{m'_{\mathbb{H},i}}.$$

10. Proof of Theorem 3.

10.1. The complex case. As already shown in §6, it suffices to prove that, if $\Sigma = (\mathbb{R}, \mathbb{C}^q, \mathcal{B}) \in \mathcal{L}^q$ is ρ -symmetric, then it allows a minimal representation, as given in Theorem 4. By Theorem 9, it allows a minimal state space representation as (16). Now consider in the i th equation of (16), x_i as a latent variable. Using the elimination of latent variables procedure as explained in the beginning of §8, we can conclude that the i th equation of (16) constrains \tilde{w}_i to satisfy an equation of the form $(\Lambda_i (d/dt) \otimes I_{n_i})\tilde{w}_i = 0$. This yields the i th representation of (7) and results in the desired representation of Theorem 4.

10.2. The real case. For the real case, we can use this elimination of latent variables procedure unchanged for each of the real and complex equations in (17). These will yield the corresponding real, respectively complex, term in (11) of Theorem 5. However, the quaternionic equations in (17) require separate attention, since the quaternions \mathbb{H} do not form a field. This brings us to the following algebraic excursion.

Let \mathcal{R} be a ring with an identity. Let $\mathcal{R}[s]$ denote the ring of polynomials with coefficients in \mathcal{R} , and $\mathcal{R}^{n_1 \times n_2}[s]$ the $n_1 \times n_2$ matrices with elements in $\mathcal{R}[s]$. Furthermore, let $GL(n, \mathcal{R}[s])$ denote the group of units of the ring $\mathcal{R}^{n \times n}[s]$, i.e., the set of polynomial matrices over \mathcal{R} with a polynomial inverse. We call these matrices unimodular. Now consider the problem of bringing a given element $P \in \mathcal{R}^{n_1 \times n_2}[s]$ into a convenient canonical form by premultiplying it by a suitable element $U_1 \in GL(n_1, \mathcal{R}[s])$ and postmultiplying it by a suitable element $U_2 \in GL(n_2, \mathcal{R}[s])$. We will say that P can be brought in *diagonal form* if there exist such U_1, U_2 yielding for $U_1 P U_2$ a matrix of the form

$$U_1 P U_2 = \begin{bmatrix} D & \vdots & 0 \\ \cdots & \ddots & \cdots \\ 0 & \vdots & 0 \end{bmatrix},$$

with D a diagonal polynomial matrix. It is well known that any P can be brought in diagonal form if $\mathcal{R} = \mathbb{R}$ or \mathbb{C} .

For what rings \mathcal{R} such a diagonalization is possible? We now show that it is sufficient for \mathcal{R} to be a division ring. Recall that a ring is a *division ring* (also called a *skew field*) if its nonzero elements form a group (\mathcal{R} must contain an identity distinct from the zero, and $a \neq 0$ must imply that it is a unit of \mathcal{R} (i.e., it has an inverse $a^{-1} \in \mathcal{R}$)).

LEMMA 11. *Let \mathcal{R} be a division ring. Then each $P \in \mathcal{R}^{n_1 \times n_2}[s]$ can be brought in diagonal form.*

Proof. The proof is an adaption of the case of a skew field of the proof of the Smith form for real or complex polynomial matrices as given in [3].

If $P = 0$, there is nothing to prove. Otherwise, let P_{kl} be the element of P of least degree. By permuting rows and columns we can assume that this element is in $(1, 1)$ entry. Now assume that the $(1, 2)$ entry is also nonzero. Then divide P_{12} by P_{11} with remainder, yielding $P_{12} = P_{11}d + r$ with degree $r < \text{degree } P_{11}$. Now postmultiply with the unimodular matrix

$$\begin{bmatrix} 1 & -d & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

This replaces the element P_{12} by r . If $r \neq 0$, move it to the $(1, 1)$ position. Repeat this process for each element of the first row and first column (using the division with remainder $P_{21} = dP_{11} + r$ and premultiplication). Each time the division is carried out with a nonzero remainder, the degree of the $(1, 1)$ element must decrease. Hence after a finite number of steps, we will obtain a matrix of the form

$$\begin{bmatrix} * & 0 & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix},$$

and we obtain the lemma by induction. \square

Now consider the ring \mathbb{H} of quaternions over \mathbb{R} : expressions of the type $\alpha + i\beta + j\gamma + k\delta$ with $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ and i, j, k elements satisfying the multiplication rules as in Example 8. With the obvious rules of addition and multiplication, \mathbb{H} is a ring. However, \mathbb{H} is not commutative, but it is easy to see that it is a division ring. Hence a polynomial matrix $P \in \mathbb{H}^{n_1 \times n_2}[s]$ can be brought in diagonal form.

Now consider a system of differential equations.

$$(18) \quad R \left(\frac{d}{dt} \right) w = M \left(\frac{d}{dt} \right) \ell,$$

where $w \in C^\infty(\mathbb{R}; \mathbb{H}^q)$, $\ell \in C^\infty(\mathbb{R}; \mathbb{H}^d)$, $R \in \mathbb{H}^{f \times q}[s]$, and $M \in \mathbb{H}^{f \times d}[s]$. Note that this is a differential equation with latent variables as (12) but in which the signals and coefficients take their values in the skew field of quaternions \mathbb{H} . We would like to eliminate the latent variables in (19), using only operations in \mathbb{H} .

Observe that the manifest behavior of (19) remains invariant if we replace R with $U_1 R$ and M with $U_1 M U_2$ with U_1 and U_2 both unimodular polynomial matrices with coefficients in \mathbb{H} . By Lemma 11, U_1 and U_2 can be chosen such that

$$U_1 M U_2 = \begin{bmatrix} 0 \\ \cdot \\ D \end{bmatrix}$$

with $D = \text{diag}(d_1, d_2, \dots, d_r)$ and $0 \neq d_i \in \mathbb{H}[s]$. Denote the conformable partition of $U_1 R$ by

$$U_1 R = \begin{bmatrix} R' \\ \cdot \\ \ddot{R}'' \end{bmatrix}.$$

Now observe that, if $0 \neq d \in \mathbb{H}[s]$, then the operator $d(d/dt) : C^\infty(\mathbb{R}; \mathbb{H}) \rightarrow C^\infty(\mathbb{R}; \mathbb{H})$ is surjective. To see this, it suffices to write this as a differential operator from $C^\infty(\mathbb{R}; \mathbb{R}^4)$ into itself and use the fact that the corresponding (4×4) polynomial matrix with real coefficients has a nonzero determinant.

This implies that the manifest behavior of (19) is governed by

$$(19) \quad R' \left(\frac{d}{dt} \right) w = 0.$$

Applying this elimination result to each of the quaternionic equations in (17) yields the corresponding quaternionic equations in (11).

This establishes that Theorem 10 yields the desired representation of Theorem 5.

11. Group representations in $GL(n, \mathbb{K}[s])$. Theorem 3 implies an interesting result about abstract group representations. A mapping $\rho : G \rightarrow GL(n, \mathbb{K}[s])$ (the unimodular $n \times n$ polynomial matrices over \mathbb{K}), which is a group homomorphism and continuous (in the sense that the map $g \in G \mapsto \rho_g(\lambda) \in GL(\mathbb{K}^n)$ is continuous for each fixed $\lambda \in \mathbb{K}$ and the map $\lambda \in \mathbb{K} \mapsto \rho_g(\lambda) \in GL(\mathbb{K}^n)$, is continuous uniformly in $g \in G$), is called a *representation of G on $GL(n, \mathbb{K}[s])$* . Two such representations ρ_1 and ρ_2 are said to be *isomorphic* if there exist a $U \in GL(n, \mathbb{K}[s])$ such that $\rho_{2,g}(s) = U(s)\rho_{1,g}(s)(U(s))^{-1}$. The representation ρ is said to be a *constant representation* if $\rho_g(s)$ is actually a constant matrix for all $g \in G$, that is, if $\rho : G \rightarrow GL(\mathbb{K}^n)$, if ρ is actually a representation of G on \mathbb{K}^n . We now show that Theorem 3 implies that *every* representation of G on $GL(n, \mathbb{K}[s])$ is isomorphic to a *constant* one. This result can be deduced from [7], where a more general theorem is proved using very different methods. Our proof, which, as we have seen, relies on the state space representation of dynamical systems in \mathcal{L}^q , provides a nice alternative “*system theoretic*” proof of this result in the theory of algebraic groups. On the other hand, we also note that Theorem 3 can be easily deduced directly from Corollary 12 (and hence from [7]). Hence Theorem 3 and Corollary 12 are fully equivalent.

COROLLARY 12. *Let $GL(n, \mathbb{K}[s])$ denote the group of unimodular $n \times n$ polynomial matrices over \mathbb{K} ($= \mathbb{R}$ or \mathbb{C}). Let G be a compact group and let $\rho : G \rightarrow GL(n, \mathbb{K}[s])$ be a representation of G on $GL(n, \mathbb{K}[s])$. Then ρ is isomorphic to a constant representation.*

Proof. 1) Consider, for each $\lambda \in \mathbb{K}$, the mapping $\rho^\lambda : G \rightarrow GL(\mathbb{K}^n)$ defined by $\rho_g^\lambda := \rho_g(\lambda)$. It is easy to see that ρ^λ is a representation of G on \mathbb{K}^n . We first prove that all the representations ρ^λ are isomorphic. It is well known that two representations of G on \mathbb{K}^n are isomorphic if and only if their characters are equal. Hence it suffices to prove that the characters $X_{\rho^\lambda} : G \rightarrow \mathbb{K}$, defined by $X_{\rho^\lambda}(g) := \text{Trace}(\rho_g(\lambda))$, are independent of $\lambda \in \mathbb{K}$. Let $\rho^\lambda \cong m_1^\lambda \rho_1 \oplus m_2^\lambda \rho_2 \oplus \cdots \oplus m_k^\lambda \rho_k \oplus \cdots$ be a decomposition of ρ^λ in terms of irreducible representations. We prove that m_k^λ is independent of λ . Now m_k^λ is given by $m_k^\lambda = \langle \chi_{\rho^\lambda}, \chi_{\rho_k} \rangle := \int_G \chi_{\rho^\lambda} \chi_{\rho_k}^* dg$, where dg is the normalized Haar measure [9] of G . This implies that m_k^λ is continuous as a function of λ for $\lambda \in \mathbb{K}$. However, since m_k^λ is integer-valued, m_k^λ must therefore be constant for $\lambda \in \mathbb{K}$. This shows that all the ρ^λ 's are isomorphic.

2) The first part of the proof yields the existence of a map $M : \mathbb{K} \rightarrow GL(\mathbb{K}^n)$ such that

$$(20) \quad \rho_g(\lambda)M(\lambda) = M(\lambda)\rho_g(0)$$

for all $\lambda \in \mathbb{K}$ and $g \in G$. We now show that this implies the existence of a polynomial matrix $R \in \mathbb{K}^{n \times n}[s]$ with $\det R \neq 0$ such that

$$(21) \quad \rho_g(s)R(s) = R(s)\rho_g(0)$$

for all $g \in G$. In other words, we show that the set of equations

$$(22) \quad \rho_g(s)X(s) = X(s)\rho_g(0), \quad g \in G$$

has a polynomial solution $X \in \mathbb{K}^{n \times n}[s]$ with $\det X \neq 0$. Note that we can rewrite (22) in vector-matrix notation as

$$(23) \quad A_g(s)x(s) = 0, \quad g \in G$$

with $x \in \mathbb{K}^{n^2}[s]$ and $A_g \in \mathbb{K}^{n^2 \times n^2}[s]$. The vector x corresponds to the elements of X in (22), while the matrix A_g corresponds to the coefficients of the linear equations in (22). Since (23) must be satisfied for each $g \in G$, there are, in principle, an infinite number of linear equations in (23). Now consider the rows of the matrices A_g and view them as vectors of rational functions, as elements of $\mathbb{K}^{1 \times n^2}(s)$. Write (23) as

$$\begin{bmatrix} A'(s) \\ \vdots \\ A''(s) \end{bmatrix} x(s) = 0$$

with A' such that its rows form a basis over $\mathbb{K}(\sim)$ for the span of the rows of all the matrices A_g , $g \in G$. Now observe that, by premultiplying x and A' by a unimodular matrix, we may as well assume that A' is in Smith form, as follows:

$$A'(s) = [D(s) \vdots 0]$$

with $D(s) = \text{diag}(d_1(s), d_2(s), \dots, d_k(s))$ and $d_i \neq 0$ for $i = 1, 2, \dots, k$. This yields that in a conformable partition A'' will be of the form $A''(s) = [\tilde{A}''(s) \vdots 0]$. Obviously, this implies that each vector polynomial

$$x(s) = \text{col}(x_1(s), \dots, x_k(s), x_{k+1}(s), \dots, x_{n^2}(s))$$

with $x_1 = x_2 = \dots = x_k = 0$ yields a solution $X(s)$ of (23).

To show that there is at least one of these solutions with $\det X \neq 0$, we will use (20). This equation yields, for each $\lambda \in \mathbb{K}$, a solution $m(\lambda)$ to (23) with, in (23), s replaced by λ . Pick a $\tilde{\lambda} \in \mathbb{K}$ such that $d_i(\tilde{\lambda}) \neq 0$ for $i = 1, 2, \dots, k$. Clearly, $m(\tilde{\lambda})$ must be of the form $\text{col}(0, \dots, 0, \tilde{m}_{k+1}, \dots, \tilde{m}_{n^2})$. Now pick a solution $x(s)$ of (23) such that $x(\tilde{\lambda}) = m(\tilde{\lambda})$. The solution $X(s)$ of (22) thus obtained will have $X(\tilde{\lambda}) = M(\tilde{\lambda})$, whence $\det X \neq 0$. This shows that (21) has a solution with $\det R \neq 0$.

3) Now consider the dynamical system $\Sigma = (\mathbb{R}, \mathbb{K}^n, \ker R(d/dt))$. Obviously, by (21), Σ is $\rho(0)$ -symmetric, and $R(d/dt)w = 0$ is a minimal representation of Σ . By Theorem 3, there exists a $U \in GL(n, \mathbb{K}[s])$ and a representation $\rho' : G \rightarrow GL(\mathbb{K}^n)$ such that

$$(24) \quad \rho'_g U(s) R(s) = U(s) R(s) \rho_g(0)$$

for all $g \in G$. Comparing (21) and (24) yields

$$\rho_g(s) = (U(s))^{-1} \rho'_g U(s),$$

which is the claim of the corollary. \square

Remark. Let $\rho : G \rightarrow GL(q, \mathbb{K}[s])$ be a representation of G on $GL(q, \mathbb{K}[s])$. Let $\Sigma = (\mathbb{R}, \mathbb{K}^q, \mathcal{B}) \in \mathcal{L}^q$ be ρ -symmetric, meaning that $\rho_g(d/dt)\mathcal{B} = \mathcal{B}$ for all $g \in G$. Note that this is a dynamic symmetry, in contrast (see §3.3) with the static symmetries studied in this paper. However, Corollary 12 shows that, by a dynamic change of variables $w \mapsto U(d/dt)w$, with U a suitable unimodular polynomial matrix, the study of this type of dynamic symmetry reduces to a static symmetry.

REFERENCES

- [1] R. W. BROCKETT AND J. L. WILLEMS, *Discreted partial differential equations: Examples of control systems defined on modules*, Automatica, 10 (1974), pp. 507–515.
- [2] F. FAGNANI AND J. C. WILLEMS, *Representations of time-reversible systems*, J. Math. Systems Estimation and Control, 1 (1991), pp. 5–28.
- [3] P. A. FURHMANN, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.
- [4] J. W. GRIZZLE AND S. I. MARCUS, *The structure of nonlinear control systems possessing symmetries*, IEEE Trans. Automat. Control, 30 (1985), pp. 248–258.
- [5] M. HAZEWINDEL AND C. F. MARTIN, *Symmetric linear systems: An application of algebraic systems theory*, Internat. J. Control, 37 (1983), pp. 1371–1384.
- [6] ———, *On decentralization, symmetry, and special structure in linear systems*, Lecture Notes in Control and Inform. Sci., 58 (1984), pp. 437–440.
- [7] V. G. KAC AND D. H. PETERSON, *On geometric invariant theory for infinite-dimensional groups*, Lecture Notes in Math., 1271 (1987), pp. 109–142.
- [8] C. F. MARTIN, *Linear decentralized systems with special structure*, Internat. J. Control, 35 (1982), pp. 291–308.
- [9] J.-P. SERRE, *Linear Representations of Finite Groups*, Springer-Verlag, 1977, Berlin, New York, translation of *Représentations Linéaires des Groupes Finis*, Hermann, 1971.
- [10] J. M. SCHUMACHER, *Transformations of linear systems under external equivalence*, J. Linear Algebra Its Appl., 102 (1988), pp. 1–34.
- [11] A. J. VAN DER SCHAFT, *Symmetries, conservation laws, and time reversibility for Hamiltonian systems with external forces*, J. Math. Phys., 24 (1983), pp. 2095–2101.
- [12] ———, *System Theoretic Descriptions of Physical Systems*, CWI Tract No. 3, 1984.
- [13] ———, *Symmetries in optimal control*, SIAM J. Control and Optim., 25 (1987), pp. 245–259.
- [14] J. C. WILLEMS, *Symmetries in dynamical systems*, in Proc. 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 520–521.
- [15] ———, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. on Automat. Control, 36 (1991), pp. 259–294.
- [16] ———, *Models for dynamics*, Dynam. Report. Ser. Dynam. Syst. Appl., 2 (1989), pp. 171–269.
- [17] D. C. YOULA AND P. TISSI, *N-port synthesis via reactance extraction*, IEEE Internat. Convention Record, 1966, pp. 183–205.

LANGUAGE STABILITY AND STABILIZABILITY OF DISCRETE EVENT DYNAMICAL SYSTEMS*

RATNESH KUMAR[†], VIJAY GARG[‡], AND STEVEN I. MARCUS[§]

Abstract. This paper studies the stability and stabilizability of discrete event dynamical systems (DEDSs) modeled by state machines. Stability and stabilizability are defined in terms of the behavior of the DEDSs, i.e. the language generated by the state machines (SMs). This generalizes earlier work where they were defined in terms of legal and illegal states rather than strings. The notion of reversal of languages is used to obtain algorithms for determining the stability and stabilizability of a given system. The notion of stability is then generalized to define the stability of infinite or sequential behavior of a DEDS modeled by a Büchi automaton. The relationship between the stability of finite and stability of infinite behavior is obtained and a test for stability of infinite behavior is obtained in terms of the test for stability of finite behavior. An algorithm of linear complexity for computing the regions of attraction is presented, which is used for determining the stability and stabilizability of a given system defined in terms of legal states. This algorithm is then used to obtain efficient tests for checking sufficient conditions for language stability and stabilizability.

Key words. discrete event dynamical systems, automata theory, supervisory control, stability, stabilizability

AMS subject classification. 93

1. Introduction. Ramadge and Wonham in their work [21] on supervisory control of discrete event dynamical systems (DEDS) have modeled a DEDS, also called a plant, by a state machine (SM), the event set of which is finite and is partitioned into sets of controllable and uncontrollable events. The language generated by such a SM is used as a model to describe the behavior of the plant at the logical level. The control task is to synthesize a controller, also called a supervisor, which disables some of the controllable events in the plant so that the closed-loop behavior equals some prespecified desired behavior, also called legal behavior. Supervisors that do not prevent any uncontrollable events from occurring are called complete. Thus there may not always exist a complete supervisor so that the closed loop system has a prespecified desired behavior. Attention is then restricted to designing a complete supervisor that is *minimally restrictive* [21], [20], [10], [1], [11] so that the closed-loop system can engage in some maximal behavior and still maintain the prescribed behavioral constraint. Thus the control objective is usually described as the synthesis of a minimally restrictive supervisor so that the controlled system has a maximally permissive legal behavior.

Sometimes such a constraint on the system behavior leads to the design of a supervisor which results in a very restrictive behavior [14], [15]. Recently there has been work [14], [15] on posing a supervisory control problem that allows the system to engage in some illegal behavior that can be tolerated. In this paper, we also allow the

* Received by the editors July 23, 1991; accepted for publication April 24, 1992. This research was supported in part by the Center for Robotics and Manufacturing, University of Kentucky, in part by the National Science Foundation under grants NSFD-CDR-8803012 and NSF-CCR-9110605, in part by the Air Force Office of Scientific Research (AFOSR) under contract F49620-92-J-0045, in part by a University Research Institute Grant, and in part by a Bureau of Engineering Research Grant.

[†] Department of Electrical Engineering, University of Kentucky, Lexington, Kentucky 40506.

[‡] Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Texas 78712-1084.

[§] Department of Electrical Engineering and Institute of System Research, University of Maryland, College Park, Maryland 20742.

possibility of the system behaving illegally. The supervisor is synthesized so that the behavior of the supervised system is “asymptotically legal.” In other words, the system is initially allowed to make illegal transitions, but after a finite number of transitions, the supervised system makes only legal transitions. With the above motivation, we define the stability and stabilizability of DEDSs in terms of their legal behavior.

In [16], [18], [2], [4], [3] the notion of stability and stabilizability of DEDS's has been presented in terms of the legal and illegal states of the system. In [2], [3] a stable system is one that starts from any arbitrary initial state and, after finitely many transitions, goes to one of the legal states and stays there; a stabilizable system is one for which there exists a supervisor so that the supervised system is stable. In [18] a system is said to be stable if after starting from any arbitrary initial state it visits the legal subset of states infinitely often; a system that can be made stable in the above context by the synthesis of an appropriate supervisor is called stabilizable. We define a system to be language-stable¹ if its eventual behavior remains confined to the legal behavior; if a supervisor exists such that the supervised system is language-stable, then the system is called language-stabilizable. We show below that the existence of an eventually reachable legal set of states implies the existence of an eventually reachable legal behavior, whereas the converse is not always true. Thus the notion of stability presented here is finer than those in [18], [2], [3] in the sense that there need not exist any fixed set of legal states. A state can eventually be reached by legal as well as illegal strings, so none of the states can be predefined to be legal. To illustrate this point consider, for example, an elevator which moves between three floors—bottom, middle, and top. Assume that a passenger requests service at the top floor. We can view the top floor to be the legal state and require that the elevator should eventually reach it. However, a “finer” constraint that the elevator should reach the top floor in no more than two moves (there are total three floors) may be desired. In this case the top floor may be reached by legal as well as illegal sequences of moves of the elevator. Thus in this example the stability based on legal states cannot capture the desired behavior of the elevator.

In [18], [2], [3], the supervisors considered for stabilizing a system are assumed to be of *static feedback* type in which the next control action is determined by the current state of the system. In general a supervisor can be of *dynamic feedback* type where the next control action is determined not necessarily by the current state but by the “path” taken to reach the current state. We refine the notion of stability and stabilizability by defining it in terms of languages rather than states and show that in some cases a static feedback-type supervisor cannot stabilize a system and a more general dynamic feedback type supervisor is needed for stabilization. In [16], the stability of systems under partial observation is studied. In this case, the supervisor is of dynamic feedback type; it can be represented as a cascade of a dynamic state observer followed by a static feedback-type controller. The supervisor considered for eventually restrictable systems in [17] is also of dynamic feedback type.

We start with the description of DEDSs and present some of the notions of stability defined in terms of states. The computational complexity of the algorithms presented in [2], [3] for determining the stability and stabilizability of DEDSs based on computing the *regions of attraction* is quadratic in the number of states of the system. We present an algorithm that is linear in the number of states of the system and is thus computationally more efficient. We then introduce the notion of

¹ We use the term language-stability to emphasize the fact that it is defined in terms of legal behavior rather than legal states, in which case it may be called state-stability.

stability in terms of languages and provide algorithms for determining the stability and stabilizability of a given system by considering an equivalent problem defined in terms of *reversal* of languages. We also discuss the computational complexity of these algorithms. Later, we provide computationally more efficient algorithms for testing the sufficiency of stability and stabilizability of systems based on our algorithm for computing the regions of attraction. In all this, we assume that perfect observation of the system behavior is possible so that the control actions are determined on the basis of observing the system evolution perfectly. We also introduce a weaker notion of language stability that is preserved under union and provide a technique for constructing the minimally restrictive stabilizing supervisor in this weaker sense of language stability.

The notion of language stability is then generalized to study the stability of sequential behaviors of DEDSs modeled by Büchi automata. The notions of ω -stability and ω -stabilizability are introduced in this context, and tests for verifying stability and stabilizability of sequential behavior are obtained by reducing the problem of testing them to the problem of testing language stability. We introduce an equivalence relation on the space of infinite strings and obtain a necessary condition of ω -stability in terms of this equivalence relation.

2. Notation and terminology. A DEDS to be controlled, called a *plant*, is modeled as a deterministic trim [8] state machine (SM) following the framework of [21]. Let the quintuple

$$P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m)$$

denote an SM representing a plant, where X denotes the state set; Σ denotes the finite event or alphabet set; $\alpha : \Sigma \times X \rightarrow X$ denotes the partial state transition function; $x_0 \in X$ denotes the initial state; and $X_m \subseteq X$ denotes the set of marked states. The transition function $\alpha(\cdot, \cdot)$ is extended to $\Sigma^* \times X$ in the natural way, where Σ^* denotes the set of all finite sequences of events belonging to Σ . The notation $\epsilon \in \Sigma^*$ is used to denote the empty string. The behavior of P is described by the language $L(P) \subseteq \Sigma^*$ that it *generates* and $L_m(P) \subseteq L(P)$ that it *marks* or *recognizes*. Formally,

$$L(P) = \{s \in \Sigma^* \mid \alpha(s, x_0)!\}; L_m(P) = \{s \in L(P) \mid \alpha(s, x_0) \in X_m\},$$

where the notation “!” is used to denote “is defined.” By definition, $L(P)$ is prefix-closed and also, since P is trim, $\overline{L_m(P)} = L(P)$ [8].

The event set is partitioned into $\Sigma = \Sigma_u \cup \Sigma_c$, the set of *uncontrollable* and *controllable* events. A supervisor S for controlling a plant is another DEDS, also represented as an SM,

$$S \stackrel{\text{def}}{=} (Y, \Sigma, \beta, y_0, Y_m).$$

S operates synchronously with P , thus allowing only the synchronous transitions to occur in the closed-loop system described by the SM [10], [11]

$$P \square S \stackrel{\text{def}}{=} (Z, \Sigma, \gamma, z_0, Z_m).$$

where $z_0 = (x_0, y_0)$; for $s \in \Sigma^*$, $\gamma : \Sigma^* \times Z \rightarrow Z$ is defined as $\gamma(s, z_0) = (\alpha(s, x_0), \beta(s, y_0))$ if $\alpha(s, x_0)!$ and $\beta(s, y_0)!$, undefined otherwise; $Z = \{z \in X \times Y \mid \exists s \in \Sigma^* \text{ s.t. } \gamma(s, z_0) = z\}$; and $Z_m = Z \cap (X_m \times Y_m)$. Thus $Z \subseteq X \times Y$ is the set of states that are reachable

from the initial state z_0 , and $Z_m \subseteq Z$ is the set of those reachable states that have both their “coordinates” marked.

The following remark describes the control achieved by the synchronous operation of P and S .

Remark 2.1 ([11], [10]). Let $L(P \square S)$ be the language generated and $L_m(P \square S)$ the language marked by $P \square S$; then $L(P \square S) = L(P) \cap L(S)$, and $L_m(P \square S) = L_m(P) \cap L_m(S)$, where $L(S), L_m(S)$ denote the languages generated, recognized by S , respectively.

Also, since S can disallow only the controllable events from occurring, $L(P) \cap \Sigma_u^* \subseteq L(P \square S)$, where Σ_u^* is the set of finite sequences of events belonging to Σ_u .

The supervisor as defined above represents a closed-loop control policy. This differs from open-loop control policy in which control actions are all prespecified; in closed-loop control, control actions are determined by observing all or part of the history of the system evolution.

DEFINITION 2.2. Let the map $f : L(P) \rightarrow 2^\Sigma$ denote a *control policy* as described in [21], i.e., for each string $s \in L(P)$ generated by the plant P , $f(s) \subseteq \Sigma$ is the set of events that are not disabled by a supervisor. Then the control exercised by the synchronous operation of a supervisor and the plant, as described above, defines the following control policy over the set of strings generated by the plant:

$$f(s) \stackrel{\text{def}}{=} \begin{cases} \{\sigma \in \Sigma \mid \gamma(s\sigma, z_0)!\} & \text{if } \gamma(s, z_0)! \\ \text{undefined} & \text{otherwise,} \end{cases}$$

where the string $s \in L(P)$.

Closed-loop controllers can further be classified into static and dynamic control type. Given a deterministic SM, $V \stackrel{\text{def}}{=} (Q, \Sigma, \delta, q_0, Q_m)$, there is a natural equivalence relation R_V [8], [6], [11], [10] induced by V on Σ^* , which is defined by $s \cong t(R_V) \Leftrightarrow \delta(s, q_0) = \delta(t, q_0)$ (this is meant to include the condition that $\delta(s, q_0)$ is undefined $\Leftrightarrow \delta(t, q_0)$ is undefined), where $s, t \in \Sigma^*$. Thus all those strings, which upon execution result in the same state in V , belong to the same equivalence class. We use $[s](R_V)$ to denote the equivalence class under the equivalence relation R_V containing the string s .

DEFINITION 2.3. Consider the control policy $f : L(P) \rightarrow 2^\Sigma$ defined by the synchronous composition operator as described in Definition 2.2. We say that a closed-loop control policy is *static* if $s \cong t(R_P) \Rightarrow f(s) = f(t)$ whenever both $f(s), f(t)$ are defined.

In other words, in a static feedback-type control, the same control action is applied after the execution of all strings that lead to the same state in the plant. Next we show that if a supervisor exercises a static closed-loop control, then it can be represented as an SM having structure similar to that of the plant.

DEFINITION 2.4. Let $V_1 \stackrel{\text{def}}{=} (Q_1, \Sigma, \delta_1, q_{01}, Q_{m1})$ and $V_2 \stackrel{\text{def}}{=} (Q_2, \Sigma, \delta_2, q_{02}, Q_{m2})$ be two SMs. V_1 is said to be a *subautomaton* [5] of V_2 if there exists a one-to-one map $h : Q_1 \rightarrow Q_2$ such that $h(\delta_1(s, q_{01})) = \delta_2(s, q_{02})$ for each $s \in L(V_1)$.

Thus if V_1 is a subautomaton of V_2 , then $L(V_1) \subseteq L(V_2)$. Note that if the map h in Definition 2.3 is also onto, then V_1 and V_2 are structurally identical.

PROPOSITION 2.5 ([9]). *The following are true:*

1. *If S is a subautomaton of P , then the control policy $f : L(P) \rightarrow 2^\Sigma$ defined by S is static.*
2. *If $f : L(P) \rightarrow 2^\Sigma$ is a static control policy, then there exists S that defines the same control policy as f and is a subautomaton of P .*

DEFINITION 2.6. A closed-loop control policy is said to be *dynamic* if it is not static.

Example 2.7. Consider, for example, a plant P , with language $L(P) = (a + b)^*$ defined over the event set $\Sigma = \{a, b\}$. Assume that $\Sigma_c = \Sigma$ (see Fig. 1; “○” denotes the states, an entering arrow “ \rightarrow ” to “○” represents the initial state, and “○” denotes the marked states). Then the language generated by the coupled system under

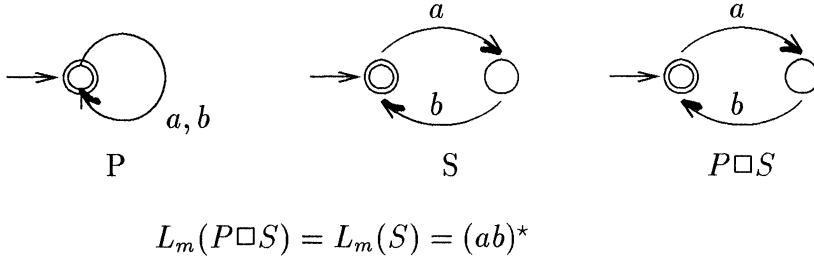


FIG. 1. Diagram illustrating Example 2.7

a static feedback control policy could be one of the following: $L(P \square S) = (a + b)^*$ or a^* or b^* or ϵ , depending on whether the events disabled in the only state of the system are \emptyset , $\{b\}$, $\{a\}$, or $\{a, b\}$.

On the other hand, the language marked by the coupled system can be made to be any sublanguage $K \subseteq (a + b)^*$ by using a dynamic feedback control policy. This can be done because all the events are controllable [10], [11] (pick the supervisor S , so that $L(S) = K$). An example for the case $K = (ab)^*$ is shown in Fig. 1.

3. Stability: region of attraction. With the above introduction on our supervisory control model, we next consider the stability issues for DEDSs. First, we discuss the definitions and results of some of the earlier works, in which the stability is defined in terms of a set of legal states of the system. Later, we present our own notions of stability defined in terms of legal behavior of the system.

Consider a plant $P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m)$. Let $\hat{X} \subseteq X$ be the prescribed subset of states or the legal states. The notions of *strong* and *weak attraction* [2], [3] are defined as follows: A state $x \in X$ is said to be *strongly attractable* to \hat{X} if, after starting from the state x , the system always reaches a state in the set \hat{X} after a finite number of transitions. The set of all the strongly attractable states is called the *region of strong attraction* of \hat{X} and is denoted by $\Omega(\hat{X})$. Formally, let for $s \in \Sigma^*$, $|s|$ denote the length of s , and for $X' \subseteq X$, let $|X'|$ denote the number of states in the set X' .

DEFINITION 3.1. $x \in X$ is strongly attractable to \hat{X} if for all s such that $\alpha(s, x)!$ and $|s| \geq |X - \hat{X}|$ there exists a prefix $u_s \in \Sigma^*$ of s with $|u_s| \leq |X - \hat{X}|$ so that $\alpha(u_s, x) \in \hat{X}$ [2,3].

DEFINITION 3.2. A state $x \in X$ is said to be *weakly attractable* to \hat{X} if there exists a static feedback supervisor S such that x is strongly attractable to \hat{X} in the coupled system $P \square S$. The set of all the weakly attractable states is called the *region of weak attraction* and is denoted by $\Lambda(\hat{X})$.

Clearly, $\Omega(\hat{X}) \subseteq \Lambda(\hat{X})$. If $\Omega(\hat{X}) = X$, then P is said to be *stable* with respect to \hat{X} and if $\Lambda(\hat{X}) = X$, then P is said to be *stabilizable* with respect to \hat{X} . Thus, to test whether a given system is stable (stabilizable) with respect to a given set of legal states, we must compute the region of strong (weak) attraction. The definitions of strongly and weakly attractable states are the same as those of *prestable* and *prestabilizable* states, respectively [18].

Remark 3.3. Algorithms for constructing the regions of strong and weak attraction are presented in [18], [2], [3]. The complexity of these algorithms is quadratic in number of states of the system. An algorithm of linear time complexity in the number of states of the system for constructing the regions of strong and weak attraction is presented in Appendix A of this paper. This algorithm is used later for arriving at a computationally more efficient test for determining a sufficient condition of stability and stabilizability introduced below.

4. Language-Stability. So far we have discussed stability of DEDSs defined in terms of their legal states and provided an efficient algorithm for testing it by computing the regions of attraction (refer to Appendix A for the algorithm). Next, we provide motivation for a more general notion of stability, which we call language-stability, and discuss some of the issues related to stability in this framework.

In some cases, it might be desirable that the eventual behavior (rather than the whole behavior) of the system be legal, so the whole behavior of the system need not be confined to a legal language as in [21], [20]. Thus in these cases the control task can be formulated as the synthesis of a supervisor such that the behavior of the supervised system is eventually legal. This leads to the design of supervisors that are less restrictive and as a result, the behavior of the supervised system is a larger language. Hence, we will formalize the notion of eventual behavior of the systems and define stability and stabilizability of systems in terms of their behavior. As discussed in the previous sections, the notions of stability defined in terms of languages can also be viewed as a generalization to the ones defined in terms of states [18], [16], [2], [4], and [3].

Example 4.1. Consider the machine P shown in Fig. 2. P can either be in “idle,” “working,” “broken,” or “display” state. Assume that initially it is in the idle state and goes to the working state when the action “start” is executed. While in the working state, P can either “stop” and go back to the idle state or can “fail” and go to the broken state. In the broken state it can execute either the action “repair” and go to the display state or the action “replace” and go back to the initial idle state. While in the display state, the action “reject” or “approve” can be executed, so the resulting state of P can either be broken (if reject is executed) or idle (if approve is executed).

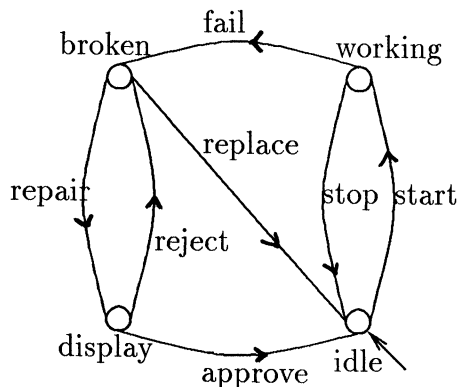


FIG. 2. Machine P of Example 4.1.

Consider the above example for the stability analysis in the framework of [18], [16], [2], [4], and [3]. The states idle and working are the “good” or legal states of P . The actions start, repair, and replace are the controllable actions, whereas the actions stop, fail, reject, and approve are the uncontrollable actions. Clearly, P is not *stable* with respect to its legal states (once P executes fail, it is not guaranteed to get back to the legal states). To show that P is *stabilizable*: once it executes fail and goes to the broken state, it must execute the controllable action replace to go back to the legal state either permanently (as in [4], [2], [3]) or temporarily (as in [18]). Suppose instead that it executes the controllable action repair and goes to the display state; there it might not execute the uncontrollable action approve in which case it would remain in the illegal state. Hence the only way P can be stabilized is by executing the action replace after it executes fail. This however, may not be desired, for replacing (and not repairing) P whenever it fails might be cost ineffective. Thus in this example, the framework of [18], [16], [3] may be too restrictive for *stabilizing* the machine P .

We would like the desired behavior of P to be such that it allows P to execute the repair-reject sequence for a finite number of times. In other words, the desired behavior of P is that if it executes fail, it should execute replace or approve after a finite number of executions of the repair-reject sequence; otherwise it should execute the start-stop sequence. The way P is designed, after executing fail, it might never execute replace or approve and continue executing the repair-reject sequence, in which case the desired behavior is not achieved. We note that the desired behavior of P as described above cannot be achieved by use of a static feedback controller.

Moreover, in the above example, P is allowed to execute “illegal” actions (the repair-reject sequence) after it executes fail, provided it eventually executes one of the “legal” actions (replace or approve). Thus the whole behavior of the system need not always be confined to a legal language as in [21], [20].

With this motivation, we formally define stability of systems in terms of their legal behavior. For $n \in \mathcal{N}$, let Σ^n denote the set of strings, each of length n , of events belonging to Σ . We use $\Sigma^{\leq N}$ to denote $\bigcup_{n \leq N} \Sigma^n$ for each $N \in \mathcal{N}$.

DEFINITION 4.2. Let $L, K \subseteq \Sigma^*$ be two languages. L is said to be *language stable* (ℓ -stable) with respect to K if there exists $N \in \mathcal{N}$ such that $L \subseteq \Sigma^{\leq N} K$.

Since $\Sigma^{\leq N} \subseteq \Sigma^{\leq N'}$ whenever $N \leq N'$ ($N, N' \in \mathcal{N}$), it follows that if L is ℓ -stable with respect to K , then there exists a smallest integer $N_0 \in \mathcal{N}$ such that $L \subseteq \Sigma^{\leq N_0} K$. Given a string $s \in \Sigma^*$, let $u_n \in \Sigma^*$ be the prefix of length n of s ($n < |s|$), and let $v_n \in \Sigma^*$ be such that $s = u_n v_n$. We define a map $\Pi_n : \Sigma^* \rightarrow \Sigma^*$ in the following manner:

$$\Pi_n(s) = \begin{cases} v_n, & \text{for } n < |s| \\ \epsilon, & \text{otherwise} \end{cases}$$

Thus the effect of the map $\Pi_n(\cdot)$ on a string s is to remove the initial n symbols of s .

It follows from Definition 4.2 that $L \subseteq \Sigma^*$ is ℓ -stable with respect to $K \subseteq \Sigma^*$ if and only if there exists $N \in \mathcal{N}$ such that for every string $s \in L$ there exists a prefix $u_s \in \Sigma^*$ of s with $|u_s| \leq N$ such that $\Pi_{|u_s|}(s) \in K$. Thus L is ℓ -stable with respect to K , if after removing a prefix of length at most N from a string in L , it matches some string in K . The language L can be thought to be representing the plant behavior and the language K can be thought to be representing the eventual legal behavior of plant. If L is not ℓ -stable with respect to K , then it is said to be *ℓ -stabilizable* with respect to K if there exists a supervisor S such that the closed loop behavior is ℓ -stable with respect to K . Formally,

DEFINITION 4.3. Consider $L, K \subseteq \Sigma^*$. L is said to be ℓ -stabilizable with respect to K if there exists a nonempty controllable [21] sublanguage $H \subseteq L$ such that H is ℓ -stable with respect to K .

Assume that L is recognized by a plant P , i.e. $L_m(P) = L$. Let S be a supervisor such that the language recognized by the closed loop system $L_m(P \square S)$ is ℓ -stable and controllable with respect to K ; then clearly L is ℓ -stabilizable with respect to K with $H = L_m(P \square S)$. It is known that the closed-loop behavior $L_m(P \square S)$ is controllable if and only if S is a *complete*² supervisor [21], [11], [10]. Thus Definition 4.3 can equivalently be stated as: L is said to be ℓ -stabilizable with respect to K if there exists a complete supervisor S such that $L_m(P \square S)$ is ℓ -stable with respect to K .

PROPOSITION 4.4. If $P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m)$ is stable (stabilizable) with respect to $\hat{X} \subseteq X$, then $L_m(P)$ is ℓ -stable (ℓ -stabilizable) with respect to $\bigcup_{x \in \hat{X}} L_m(P, x)$, where $L_m(P, x)$ is the language marked by P assuming the initial state to be x .

Proof. Assume that the SM, $P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m)$ is stable with respect to the legal set $\hat{X} \subseteq X$. Let $L = L_m(P)$, and $K = \bigcup_{x \in \hat{X}} L_m(P, x)$. Define $N \stackrel{\text{def}}{=} |X - \hat{X}|$. We will show that $L \subseteq \Sigma^{\leq N} K$. Consider $s \in L$. If $|s| \leq N$, then $s \in \Sigma^{\leq N}$; hence $s \in \Sigma^{\leq N} K$. If $|s| > N$, then there exists a prefix $u_s < s$, $|u_s| \leq N$, such that $\alpha(u_s, x_0) \in \hat{X}$ (follows from the fact that x_0 is strongly attractable to \hat{X}). Thus $\Pi_{|u_s|}(s) \in K$ (by definition of K). Hence $s \in \Sigma^{\leq N} K$; which shows that L is ℓ -stable with respect to K .

Similarly, it can be shown that if P is stabilizable with respect to \hat{X} , then L is ℓ -stabilizable with respect to K . \square

Proposition 4.4 shows that stability (stabilizability) in terms of states in some sense implies ℓ -stability (ℓ -stabilizability). We show in the next example that the converse does not necessarily hold, thus showing that the notion of ℓ -stability (ℓ -stabilizability) is finer than that of stability (stabilizability).

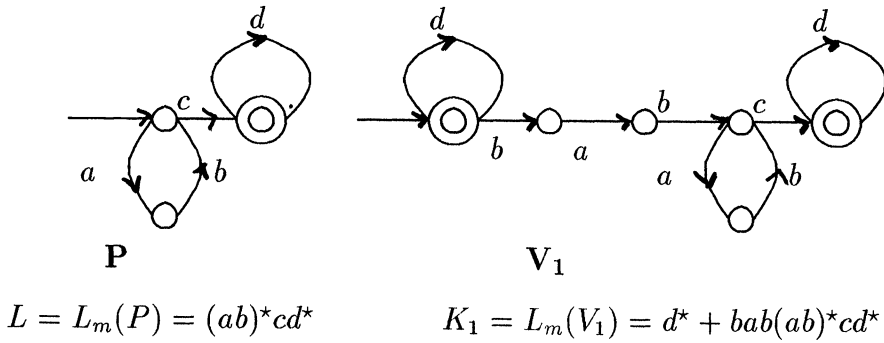
Example 4.5. Let $\Sigma = \Sigma_u = \{a, b, c, d\}$. Consider the languages $L, K_i (i \geq 1) \subseteq \Sigma^*$ given by: $L = (ab)^* cd^*$ and $K_i = d^* + b(ab)^i (ab)^* cd^*$. Generators for L and $K_1 = d^* + bab(ab)^* cd^*$ are shown in Fig. 3. Letting $N \stackrel{\text{def}}{=} 2i + 1$, it can be easily verified that $L \subseteq \Sigma^{\leq N} K_i$ for each $i \geq 1$, and also that N is the smallest integer for which the last inclusion holds. Fix, for example, $i = 1$. We show that $L \subseteq \Sigma^{\leq 3} K_1$. L consists of strings cd^* (no ab followed by cd^*), $abcd^*$ (one ab followed by cd^*), and $(ab)^{\geq 2} cd^*$ (two or more ab followed by cd^*). First, consider the strings in $cd^* \subseteq L$. Then $\Pi_1(cd^*) = d^* \subseteq K_1$. Next, consider the strings in $abcd^* \subseteq L$. Then $\Pi_3(abcd^*) = d^* \subseteq K_1$. Finally, consider $(ab)^{\geq 2} cd^* \subseteq L$. Then $\Pi_1((ab)^{\geq 2} cd^*) = b(ab)^{\geq 1} cd^* = bab(ab)^* cd^* \subseteq K_1$. Thus it follows that $L \subseteq \Sigma^{\leq 3} K_1$.

Since L is ℓ -stable with respect to each K_i it follows that L is also ℓ -stabilizable with respect to each K_i .

Let P, V_i be the minimal SMs generating L, K_i , respectively. Then P, V_i must have $3, 2i + 4$ states, respectively (refer to Fig. 3 for $i = 1$). It can be easily seen that P is not stable with respect to any of its subset of states. Since $\Sigma_u = \Sigma$, P is not stabilizable with respect to any of its subset of states either.

Example 4.6. Consider the languages $L = (ac + b)a(a + b)^*$ and $K = (ab)^*$ defined over $\Sigma = \Sigma_c = \{a, b, c\}$. We will show that L is not ℓ -stable with respect to K , i.e., there exists no $N \in \mathcal{N}$ such that $L \subseteq \Sigma^{\leq N} K$. To prove this, we assume for contradiction that there exists $N_0 \in \mathcal{N}$ is such that $L \subseteq \Sigma^{\leq N_0} K$. Consider the

² A supervisor S is said to be *complete* if for all $s \in \Sigma^*, \sigma_u \in \Sigma_u : s \in L(P \square S), s\sigma_u \in L(P) \Rightarrow s\sigma_u \in L(P \square S)$.

FIG. 3. Diagram illustrating Example 4.5 with $i = 1$.

string $baa^{N_0} \in L$. Any substring of it obtained by removing an initial finite segment of length less than N_0 does not match any string in K (a string in K contains the symbol b at the end, whereas the string baa^{N_0} ends with the symbol a).

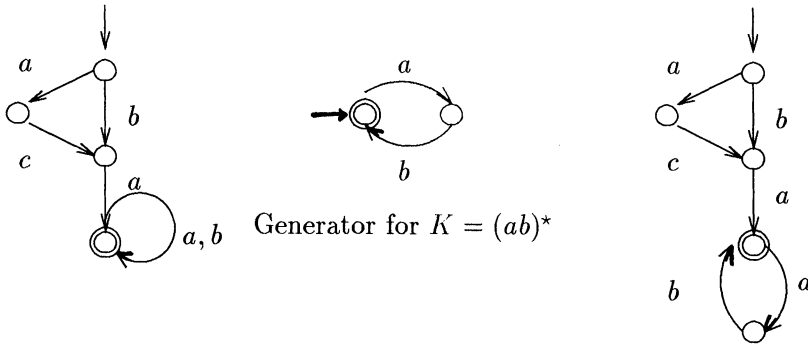


FIG. 4. Diagram illustrating Example 4.6.

Consider a sublanguage $H = (ac + b)a(ab)^* \subseteq L$ as shown in Fig. 4. Since $\Sigma_c = \Sigma$, H is controllable with respect to L . It can be easily seen that $H = (ac + b)a(ab)^*$ is ℓ -stable with respect to $K = (ab)^*$ (consider any string from H and remove the initial segment, either aca or ba , whichever is appropriate; the resulting string belongs to K). Thus L is ℓ -stabilizable to K .

In this example, it is clear that a dynamic feedback-type supervisor has been used to ℓ -stabilize the given language. Also, a static feedback-type control cannot be used to stabilize $L = (ac + b)a(a + b)^*$ with respect to $K = (ab)^*$. This follows since any string in K contains an equal number of a 's and b 's, and L cannot be restricted to a language $H \subseteq L$ with all its strings having an equal number of a 's and b 's at its end by using a static supervisor (refer to Example 2.7). In [18], [2], [3], where stability is defined in terms of the legal states, the supervisors considered for stabilizing DEDSs are all assumed to be of static feedback type. Thus a more general type of control is needed to ℓ -stabilize the behavior of a given system, which also shows that the notion of ℓ -stability (ℓ -stabilizability) is a finer notion.

Example 4.7. Consider a system consisting of a single buffer of unbounded capacity. Only two types of events arrival, denoted a , and departure, denoted b , occur in this system, i.e., $\Sigma = \{a, b\}$. The behavior of this system can be described by the language

$$L = \{s \in \Sigma^* \mid \#(a, s) \geq \#(b, s)\},$$

where the symbol $\#(x, y)$ is used to denote the number of times the symbol x occurs in the string y . We may be interested in determining whether there exists some number $N \in \mathcal{N}$ such that after execution of all strings of length larger than N , the buffer content is bounded above by a fixed number $N_0 \in \mathcal{N}$. The above problem can be posed as a ℓ -stability problem with the “eventually reachable” language $K \subseteq \Sigma^*$ defined as

$$K = \{s \in \Sigma^* \mid \#(a, s) - \#(b, s) \leq N_0\}.$$

K corresponds to the content of the buffer being bounded above by N_0 .

It is easy to see that L is not ℓ -stable with respect to K , i.e., there does not exist any N such that after execution of all strings of length larger than N , the buffer content is bounded above by N_0 . Note that in this example, $K \subseteq L$ and if the arrival event a is controllable, then L can be restricted to the language K by disabling a whenever the buffer content becomes equal to N_0 . This proves that L is ℓ -stabilizable with respect to K . Next we present algorithms for testing ℓ -stability and ℓ -stabilizability of a language L with respect to another language K .

4.1. Algorithms for testing ℓ -stability and ℓ -stabilizability. To test whether a language L is ℓ -stable (ℓ -stabilizable) with respect to another language K , we must test whether there exists an integer $N \in \mathcal{N}$ such that $L \subseteq \Sigma^{\leq N} K$ ($H \subseteq \Sigma^{\leq N} K$, where $H \subseteq L$). This problem can equivalently be posed in terms of the *reversal* [1] of languages that we define next.

DEFINITION 4.8. Given a string $s \in \Sigma^*$, its *reversal* $s^R \in \Sigma^*$, is the string obtained by reversing s . Given a language $L \subseteq \Sigma^*$, its reversal $L^R \subseteq \Sigma^*$ is defined to be: $L^R \stackrel{\text{def}}{=} \{s^R \in \Sigma^* \mid s \in L\}$.

Next we discuss some of the properties of the reversal operator. We use L, L_1, L_2 to denote languages defined on Σ .

LEMMA 4.9.

1. *Reversal preserves regularity, i.e., if L is regular, then so is L^R .*
2. $(L^R)^R = L$.
3. *Reversal is monotone, i.e., if $L_1 \subseteq L_2$, then $L_1^R \subseteq L_2^R$.*
4. $(L_1 L_2)^R = L_2^R L_1^R$.

Proof. 1. The proof is based on constructing an FSM that recognizes L^R using an FSM realization for L , and can be found in [8].

2. The proof follows from the definition of the reversal of languages and the fact that for any string $s \in \Sigma^*$, $(s^R)^R = s$.

3. Pick $s \in L_1^R$; then $s^R \in L_1$. Since $L_1 \subseteq L_2$, it follows that $s^R \in L_2$, i.e., $(s^R)^R = s \in L_2^R$.

4. We first show that $(L_1 L_2)^R \subseteq L_2^R L_1^R$. Pick $s \in (L_1 L_2)^R$; then $s^R \in L_1 L_2$, i.e., there exist $u_s \in L_1$ and $v_s \in L_2$ such that $u_s v_s = s^R$. Hence $s = (s^R)^R = (u_s v_s)^R = v_s^R u_s^R \in L_2^R L_1^R$.

Next we show that $L_2^R L_1^R \subseteq (L_1 L_2)^R$. Pick $s \in L_2^R L_1^R$; then there exist $v_s \in L_2$ and $u_s \in L_1$ such that $v_s^R u_s^R = s$. Hence $s = (s^R)^R = ((v_s^R u_s^R)^R)^R = (u_s v_s)^R \in (L_1 L_2)^R$. \square

COROLLARY 4.10. $L \subseteq \Sigma^{\leq N} K$ if and only if $L^R \subseteq K^R \Sigma^{\leq N}$, where $L, K \subseteq \Sigma^*$ and $N \in \mathcal{N}$.

Proof. Assume that $L \subseteq \Sigma^{\leq N} K$; then it follows from part 3 of Lemma 4.9 that $L^R \subseteq (\Sigma^{\leq N} K)^R$. Since $(\Sigma^{\leq N})^R = \Sigma^{\leq N}$, it follows from part 4 of Lemma 4.9 that $L^R \subseteq K^R \Sigma^{\leq N}$.

Assume next that $L^R \subseteq K^R \Sigma^{\leq N}$; then from part 3 of Lemma 4.9 it follows that $(L^R)^R \subseteq (K^R \Sigma^{\leq N})^R$. Thus from part 4 of Lemma 4.9 we obtain $(L^R)^R \subseteq \Sigma^{\leq N} (K^R)^R$. It then follows from part 2 of Lemma 4.9 that $L \subseteq \Sigma^{\leq N} K$. \square

Thus the problem of testing ℓ -stability of a language L with respect to another language K can be equivalently posed as that of determining an integer $N \in \mathcal{N}$, if it exists, such that $L^R \subseteq K^R \Sigma^{\leq N}$. Hence, given two languages $L, K \subseteq \Sigma^*$, we next analyze the problem of determining an integer $N \in \mathcal{N}$, if it exists, such that $L^R \subseteq K^R \Sigma^{\leq N}$.

Let $P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m)$ and $V \stackrel{\text{def}}{=} (Q, \Sigma, \delta, q_0, Q_m)$ be two SMs such that $L_m(P) = L^R$ and $L_m(V) = K^R$. Assume further that P is trim [8] so that $L(P) = \overline{L_m(P)} = \overline{L^R}$, and V is such that $L(V) = \Sigma^*$, i.e. V is an SM that recognizes K^R and has an additional dump state in order to generate Σ^* . Consider a slightly different synchronous composition of P and V , denoted $P \square' V$, given by the 5-tuple:

$$P \square' V \stackrel{\text{def}}{=} (R, \Sigma, \rho, r_0, R_m),$$

where the state set R , the transition function $\rho(\cdot, \cdot)$, and the initial state r_0 are defined as in the definition of synchronous composition in §2, and

$$R_m = \{r \in X_m \times Q \mid \exists s \in \Sigma^* \text{ s.t. } \rho(s, r_0) = r\}.$$

This is a slight variation to the earlier definition of marked states in synchronous composition of two state machines. Note that R_m consists of those states in $X_m \times Q$ that are reachable from the initial state r_0 ; hence $R_m \subseteq X_m \times Q$. Also, note that all transitions are defined in all states of V , i.e., given any event $\sigma \in \Sigma$ and any state $q \in Q$, $\delta(\sigma, q)!$. Hence for any event $\sigma \in \Sigma$ and state $r = (x, q) \in R$, $\rho(\sigma, (x, q))$ is defined if and only if $\alpha(\sigma, x)$ is defined.

LEMMA 4.11. *Let P and V be the two SMs as defined above. Then $L_m(P \square' V) = L_m(P)$, and $L(P \square' V) = L(P)$.*

Proof. First we show that $L_m(P \square' V) \subseteq L_m(P)$. Pick $s \in L_m(P \square' V)$; then $\rho(s, r_0) \in R_m$. Since $\rho(s, r_0)$ is defined if and only if $\alpha(s, x_0)$ is defined, and $R_m \subseteq X_m \times Q$, it follows that $\alpha(s, x_0) \in X_m$. Thus $s \in L_m(P)$. Next we show that $L_m(P) \subseteq L_m(P \square' V)$. Pick $s \in L_m(P)$; then $\alpha(s, x_0) \in X_m$. Again, since $\alpha(s, r_0)$ is defined if and only if $\rho(s, r_0)$ is defined, $\rho(s, r_0) \in X_m \times Q$. The state $\rho(s, r_0)$ is clearly a reachable state from r_0 , hence $\rho(s, r_0) \in R_m$, which shows that $s \in L_m(P \square' V)$.

Since $L(P \square' V) = L(P) \cap L(V) = L(P) \cap \Sigma^* = L(P)$, the other result follows. \square

Given two languages $L, K \subseteq \Sigma^*$, next we present a necessary and sufficient condition to determine whether there exists an integer $N \in \mathcal{N}$ such that $L^R \subseteq K^R \Sigma^{\leq N}$ in terms of the graphical structure of SMs recognizing the languages L^R, K^R .

Consider R , the state set of $P \square' V$. Let R^* denote the set of all finite sequences of states belonging to R . Consider $p \in R^*$ such that $p = (r_1 r_2 \dots r_i \dots r_n) \in R^*$, where $r_i \in R$ for each $1 \leq i \leq n$ and $n \in \mathcal{N}$. Then p is said to be a *path* starting at r_1 and ending at r_n in $P \square' V$, if there exist a string $s_p \in \Sigma^*$, $s_p = \sigma_1 \sigma_2 \dots \sigma_i \dots \sigma_{n-1}$, where $\sigma_i \in \Sigma$ for each $1 \leq i \leq n-1$, such that $\rho((\sigma_1 \dots \sigma_{i-1}), r_1) = r_i$ for each $1 < i \leq n$. $s_p \in \Sigma^*$ as described above is called the string corresponding to path p . Thus, given

a path p in $P \square' V$, there exists at least one string $s_p \in \Sigma^*$ corresponding to p . A state $r \in R$ is said to be a *path-state* of the path p if $r = r_i$ for some $1 \leq i \leq n$. p is said to be a *loop-path* if there exist i, j with $1 \leq i < j \leq n$ such that $r_i = r_j$, in which case the portion $r_i \dots r_j$ of p is called the *loop-portion* of p . p is said to be a *loopfree-path* if p is not a loop-path.

THEOREM 4.12. *Let $L^R, K^R \subseteq \Sigma^*$ be the languages recognized by the SMs P, V , respectively, as described above. Then there exists an integer $N \in \mathcal{N}$ such that $L^R \subseteq K^R \Sigma^{\leq N}$ if and only if the following hold in the SM $P \square' V$:*

(C1) *For each $r_m \in R_m$ and for every path p in $P \square' V$ that starts at r_0 and ends at r_m , there exists a path-state $r = (x, q) \in X \times Q$ of p such that $q \in Q_m$.*

(C2) *For each $r = (x, q) \in X \times Q_m$ and each $r_m \in R_m$, if a path p in $P \square' V$ that starts at r and ends at r_m has none of its path-states in $X \times Q_m$ (other than the one at which it starts), then p is a loopfree-path.*

Proof. Assume that there exists an integer $N \in \mathcal{N}$ such that $L^R \subseteq K^R \Sigma^{\leq N}$; then we first show that (C1) holds.

Fix a path p in $P \square' V$ such that p starts at r_0 and ends at $r_m \in R_m$. Then there exists a string $s_p \in L_m(P \square' V)$ such that $\rho(s_p, r_0) = r_m$. Since $L_m(P \square' V) = L_m(P) = L^R$ (Lemma 4.9 and definition of P), $s_p \in L^R$. Thus it follows from the assumption that $s_p \in K^R \Sigma^{\leq N}$, i.e., there exist $u_{s_p} \in K^R$ and $v_{s_p} \in \Sigma^{\leq N}$ such that $s_p = u_{s_p} v_{s_p}$. Consider the path-state $r = (x, q) = \rho(u_{s_p}, r_0)$ of p . Since $u_{s_p} \in K^R$, the state q reached by accepting u_{s_p} in V belongs to Q_m , i.e., $r = (x, q) \in X \times Q_m$.

Next we show that (C2) holds. Fix a path p in $P \square' V$ such that p starts at $r = (x, q) \in X \times Q_m$ and ends at $r_m \in R_m$ and none of the path-states of p other than the first one are in $X \times Q_m$. Assume for contradiction that (C2) is false, i.e., p is a loop-path. Consider the string $s \in L(P \square' V)$ such that $\rho(s, r_0) = r = (x, q)$. Since $q \in Q_m$, $s \in L_m(V) = K^R$. Let $t_p = u_p v_p w_p \in \Sigma^*$ be a string corresponding to the path p , where v_p represents the string corresponding to the loop-portion of p . Then $st_p = su_p v_p w_p \in L_m(P \square' V) = L^R$ (since $\rho(st_p, r_0) = r_m \in R_m$). Hence the string $tu_p(v_p)^{N+1}w_p \in L^R$. Then there exists no prefix $s' \in K^R$ of the string $su_p(v_p)^{N+1}w_p$ such that $\Pi_{|s'|}(su_p(v_p)^{N+1}w_p) \in \Sigma^{\leq N}$, which contradicts the fact that $L^R \subseteq K^R \Sigma^{\leq N}$. This completes the proof of the fact that (C1) and (C2) are necessary conditions for an integer $N \in \mathcal{N}$ to exist such that $L^R \subseteq K^R \Sigma^{\leq N}$. It remains to show that (C1) and (C2) are sufficient conditions also.

Assume then that (C1) and (C2) hold for SM $P \square' V$. Since (C2) holds, any path p in $P \square' V$ that starts at $r = (x, q) \in X \times Q_m$ and ends at $r_m \in R_m$ with none of its path-states (other than the first one) in $X \times Q_m$, is a loopfree-path. Let \mathcal{P} denote the collection of all such paths (paths that satisfy condition (C2)). Define $N \stackrel{\text{def}}{=} \max_{p \in \mathcal{P}} |p|$, where $|p|$ denotes the length of path p . Then we will show that $L^R \subseteq K^R \Sigma^{\leq N}$. Note that since (C2) holds, all the paths $p \in \mathcal{P}$ are loopfree-paths, hence the maximum in the definition of N exists. To show that $L^R \subseteq K^R \Sigma^{\leq N}$, pick $s \in L^R$. Then $s \in L_m(P \square' V)$. Let $\rho(s, r_0) = r_m \in R_m$. Consider the path p_s in $P \square' V$ corresponding to string s . Since $P \square' V$ is deterministic, p_s is unique. Also, p_s starts at r_0 and ends at $r_m \in R_m$. Hence by (C1), there exists a path-state $r = (x, q)$ of p_s such that $r = (x, q) \in X \times Q_m$. Let r' be the last such path-state of p_s , i.e., $r' \in X \times Q_m$ and all the path-states of p_s that follow r' do not belong to $X \times Q_m$. Let the portion of p_s that starts at r' and ends at r_m be denoted by p' ; then from (C2) p' is a loopfree-path, also $p' \in \mathcal{P}$. It follows from the definition of N that $|p'| \leq N$. Let $u' \in \Sigma^*$ be the prefix of s such that $\rho(u', r_0) = r'$, then $u' \in K^R$ (since $r' \in X \times Q_m$), and $\Pi_{|u'|}(s) \in \Sigma^{\leq N}$. Thus $s \in K^R \Sigma^{\leq N}$. This completes the proof of Theorem

4.12. \square

Remark 4.13. Conditions (C1) and (C2) can be tested in $P \square' V$ in the following manner:

1. Consider the state set R of $P \square' V$ and remove all the states $r = (x, q) \in R$ (and the transitions entering or leaving these states) for which $q \in Q_m$. Then for (C1) to hold, there must not exist any path connecting r_0 to any $r_m \in R_m$ in the machine obtained by removing the above states. Thus (C1) can be verified by performing a *single* reachability test on the reduced machine as described above.

2. Next fix a state $r = (x, q) \in R$ with $q \in Q_m$ and remove from $P \square' V$ all the other states $r' = (x', q') \in R$ (and the transitions entering or leaving these states) having $q' \in Q_m$. Then for (C2) to hold for this state r , any path connecting r to any $r_m \in R_m$ in the machine obtained by removing the above states must be acyclic. Repeat the above for every state $r'' = (x'', q'') \in R$ with $q'' \in Q_m$ and test for acyclicity. Since for each such state acyclicity can be tested by computing its reachability set, testing (C2) requires at most $|P \square' V|$ reachability tests to be performed, where $|P \square' V|$ denotes the number of states in $P \square' V$.

Thus it follows from above that testing (C1) and (C2) requires at most $|P \square' V| + 1$ reachability tests to be performed. Since the reachability set can be computed in $O(|P \square' V|)$ time, the complexity of testing (C1) and (C2) is of order $O(|P \square' V|^2)$. Let $m, n \in \mathcal{N}$ be the number of states in the minimal SMs recognizing L, K , respectively, then the number of states in SMs P, V recognizing L^R, K^R , respectively, is $2^m, 2^n$, respectively (reversal operation requires nondeterministic to deterministic conversion of SMs). Hence the computational complexity of testing ℓ -stability of L with respect to K is $O(2^{2(m+n)})$.

Example 4.14. Consider the languages $L, K_1 \subseteq \{a, b, c, d\}^*$ as in Example 4.5: $L = (ab)^*cd^*$ and $K_1 = d^* + bab(ab)^*cd^*$. Recognizers for L and K_1 are shown in Fig. 3. Then $L^R = d^*c(ba)^*$ and $K_1^R = d^* + d^*c(ba)^*bab = d^* + d^*cbab(ab)^*$. Let P, V be state machines such that $L_m(P) = L^R, L_m(V) = K_1^R$, and $L(P) = L_m(P), L(V) = \Sigma^*$, respectively, as in Theorem 4.12. Construct $P \square' V$ as described above. Recognizers for $P, V, P \square' V$ are shown in Fig. 5. A state (x, q) in the state set R of $P \square' V$ is marked

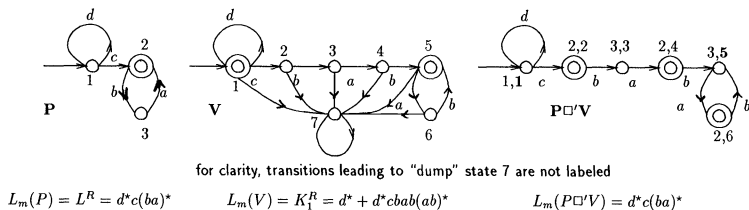


FIG. 5. Diagram illustrating Example 4.14.

if and only if the state $x \in X_m$. Since state $2 \in X$ is the only state marked in P , the states $(2, 2), (2, 4)$ and $(2, 6)$ are the marked states in $P \square' V$. We now check whether conditions (C1) and (C2) of Theorem 4.12 hold. Consider any path in $P \square' V$ starting

from the initial state $(1, 1)$ and ending at one of the marked states $(2, 2)$, $(2, 4)$, or $(2, 6)$. Then this path obviously visits the state $(1, 1)$. Since $(1, 1) \in X \times Q_m$ (state 1 is marked in V), condition (C1) holds. To show that (C2) holds, consider any path in $P \square V$ starting at the initial state $(1, 1)$ and ending at one of the marked states $(2, 2)$, $(2, 4)$, or $(2, 6)$. If the path ends at $(2, 2)$, then the last state (x, q) such that $q \in Q_m$ visited along this path is $(1, 1)$. Consider the path segment between $(1, 1)$ and $(2, 2)$; it is loopfree. If the path ends at $(2, 4)$, then the last state (x, q) with $q \in Q_m$ visited along this path is again $(1, 1)$, and again the path segment between $(1, 1)$ and $(2, 4)$ is loopfree. Finally, if the path ends at $(2, 6)$, then the last state (x, q) with $q \in Q_m$ visited is $(3, 5)$, and the path segment between $(3, 5)$ and $(2, 6)$ is again loopfree. Thus condition (C2) also holds. It then follows from Theorem 4.12 that L is ℓ -stable with respect to K_1 as expected (refer to the discussion in Example 4.5).

COROLLARY 4.15. *Consider two regular languages $L, K \subseteq \Sigma^*$. Let $m, n \in \mathcal{N}$ be the number of states in the minimal SMs recognizing L, K , respectively. L is ℓ -stable with respect to K , if and only if $L \subseteq \Sigma^{\leq 2^{m+n}} K$.*

Proof. To prove the “if” part, we must show that $L \subseteq \Sigma^{\leq 2^{m+n}} K$ implies L is ℓ -stable with respect to K . This is trivially true: set $N = 2^{m+n}$ in the definition of ℓ -stability. To prove the “only if” part, we need to show that if L is ℓ -stable with respect to K , then $L \subseteq \Sigma^{\leq 2^{m+n}} K$. Since L is ℓ -stable with respect to K , it follows from Corollary 4.10 that there exists $N \in \mathcal{N}$ such that $L^R \subseteq K^R \Sigma^{\leq N}$. Let P, V be machines recognizing L^R, K^R , respectively, as in Theorem 4.12. It then follows from Theorem 4.12 that $N \leq |P \square V|$ (refer to the second part of the proof of Theorem 4.12, where N is defined to be $N \stackrel{\text{def}}{=} \max_{p \in \mathcal{P}} |p|$; since each $p \in \mathcal{P}$ is loopfree, $|p| \leq |P \square V|$, hence $N \leq |P \square V|$). Since the number of states in SMs recognizing L, K is m, n , respectively, the number of states in P, V is $2^m, 2^n$, respectively (reversal operation requires nondeterministic to deterministic conversion of SMs [1]). Thus it follows that $N \leq |P \square V| = (2^m)(2^n) = 2^{m+n}$. \square

Remark 4.16. Thus ℓ -stability of a given language L with respect to another language K can also be determined by testing whether $L \subseteq \Sigma^{\leq 2^{m+n}} K$, where $m, n \in \mathcal{N}$ are the numbers of states present in SMs recognizing L, K , respectively.

Next we consider the problem of testing ℓ -stabilizability of a given language $L \subseteq \Sigma^*$ with respect to another language $K \subseteq \Sigma^*$. Let P be the trim SM recognizing language L , then $L(P) = \bar{L}$. The supervisor that disables all the controllable transitions of P (treated as a plant) is called the *maximally restrictive supervisor*. The behavior of P under the maximally restrictive control equals $L(P) \cap \Sigma_u^*$, which is the *infimal controllable sublanguage* of $L(P)$. The existence result and formula for the infimal controllable sublanguage of a prefix-closed language is given in [14]. Hence, if L is prefix closed, then $L(P) = \bar{L} = L$, and we obtain that the infimal controllable sublanguage of L equals $L \cap \Sigma_u^*$.

THEOREM 4.17. *If $L \subseteq \Sigma^*$ is prefix closed, then L is ℓ -stabilizable with respect to $K \subseteq \Sigma^*$ if and only if $L \cap \Sigma_u^*$ is ℓ -stable with respect to K .*

Proof. Assume that L is ℓ -stabilizable with respect to K . Then there exists $N \in \mathcal{N}$ and a nonempty controllable sublanguage $H \subseteq L$ such that $H \subseteq \Sigma^{\leq N} K$. Note that $L \cap \Sigma_u^* \subseteq H$ (by infimality of $L \cap \Sigma_u^*$). Hence $L \cap \Sigma_u^* \subseteq \Sigma^{\leq N} K$. Thus $L \cap \Sigma_u^*$ is ℓ -stable with respect to K .

Next assume that $L \cap \Sigma_u^*$ is nonempty and ℓ -stable with respect to K . Since $L \cap \Sigma_u^*$ is controllable and $L \cap \Sigma_u^* \subseteq L$, it follows that L is ℓ -stabilizable with respect to K . \square

Remark 4.18. Thus ℓ -stabilizability of a given closed language L with respect to

another language K can be determined by testing whether $L \cap \Sigma_u^*$ is nonempty and ℓ -stable with respect to K . Let m denote the number of states in the SM P that generates the language L , then $L \cap \Sigma_u^*$ can be computed in time $O(m)$ by deleting all the controllable transitions from P .

As stated in Remark 4.13, the algorithm for testing ℓ -stability of L with respect to K is of computational complexity that is exponential in the number of states present in SMs recognizing L and K . Hence so is the complexity of the algorithm that tests the ℓ -stabilizability. Next we present a sufficient condition for ℓ -stability of L with respect to K that can be tested in polynomial time. Let $P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m)$ and $V \stackrel{\text{def}}{=} (Q, \Sigma, \delta, q_0, Q_m)$ be two SMs recognizing L and K , respectively. Define the following subset of states $X_S \subseteq X$:

$$X_S = \{x \in X \mid L_m(P, x) \subseteq K\},$$

where $L_m(P, x)$ is the language recognized by P assuming its initial state to be $x \in X$.

PROPOSITION 4.19. *Consider SMs P, V as defined above. If $x_0 \in \Omega(X_S)$, then L is ℓ -stable with respect to K .*

Proof. Define $N \stackrel{\text{def}}{=} |X - X_S|$; then to prove ℓ -stability of L with respect to K , we need to show that $L \subseteq \Sigma^{\leq N} K$. Consider $s \in L$. If $|s| \leq N$, then clearly $s \in \Sigma^{\leq N} K$. So let $s \in L$ be such that $|s| > N$. Then it follows from the definition of the region of strong attraction that there exists a prefix $u_s \in \Sigma^*, |u_s| \leq N$, of s such that $\alpha(u_s, x_0) \in X_S$. Also, by the definition of X_S , $\Pi_{|u_s|}(s) \in K$, which shows that $s \in \Sigma^{\leq N} K$. \square

Thus if x_0 is strongly attractable to a state in X_S , then P , after starting from x_0 , reaches a state in X_S in at most $|X - X_S|$ transitions, and then onwards follows a string in K . The following algorithm checks the sufficient condition of ℓ -stability of Proposition 4.19.

ALGORITHM 4.20.

1. Determine the subset of states $X_S \subseteq X$ defined above.
2. Compute $\Omega(X_S)$ using Algorithm A.1.
3. If $x_0 \in \Omega(X_S)$, then L is ℓ -stable with respect to K .

Let P, V be the minimal SMs recognizing L, K , respectively, and let $m, n \in \mathcal{N}$ be the number of states in P, V , respectively. Then step 1 of Algorithm 4.20 can be determined in $O(m^2n)$ time, and steps 2 and 3 can both be determined in $O(m)$ time (refer to Theorem A.2). Hence the computational complexity of Algorithm 4.20 is $O(m^2n)$, which is polynomial in m, n . Note that Algorithm 4.20 tests only for the sufficiency condition of ℓ -stability. Hence if the condition in step 3 of Algorithm 4.20 is not satisfied, ℓ -stability of L with respect to K is determined by testing conditions (C1) and (C2) of Theorem 4.12 as described in Remark 4.13. Next we present a sufficient condition for ℓ -stabilizability of L with respect to K , which can also be tested in polynomial time.

PROPOSITION 4.21. *Consider the SMs P, V . Let $X'_S \stackrel{\text{def}}{=} \{x \in X \mid L_m(P, x) \subseteq K\}$. If $x_0 \in \Lambda(X'_S)$, then L is ℓ -stabilizable with respect to K .*

Proof. The proof is similar to the proof of Proposition 4.19. \square

The following algorithm can be used for testing the sufficient condition of ℓ -stabilizability of Proposition 4.21.

ALGORITHM 4.22.

1. Compute $X'_S \subseteq X$.

2. Compute $\Lambda(X'_S)$ using the modification to Algorithm A.1 described in Remark A.3.
3. If $x_0 \in \Lambda(X'_S)$, then L is ℓ -stabilizable with respect to K .

The computational complexity of Algorithm 4.22 is also $O(m^2n)$, where m, n is the number of states in P, V , respectively.

5. Weakly stabilizing supervisors. In the previous section we showed that given a plant P with physical behavior $L \subseteq \Sigma^*$ and desired eventual behavior $K \subseteq \Sigma^*$, it can be verified whether or not L is ℓ -stable or ℓ -stabilizable with respect to K . In the case where L is ℓ -stable with respect to K , the eventual behavior of P is contained in K ; hence no supervisor is needed. If L is not ℓ -stable but is ℓ -stabilizable with respect to K , then a supervisor must be constructed to insure that the eventual closed-loop behavior of the system is a sublanguage of K . The ℓ -stabilizability of L guarantees the existence of a stabilizing supervisor, but a minimally restrictive stabilizing supervisor need not in general exist. This is evident from the following proposition.

PROPOSITION 5.1. *ℓ -stability is not preserved under union.*

Proof. We show by the following example that ℓ -stabilizability is not preserved under union. Let $\Sigma = \Sigma_c = \{a, b\}$, $L = a^*b^*$ denote the plant behavior and $K = b^*$ denote the desired eventual behavior. Then there does not exist any integer $N \in \mathcal{N}$ such that $L \subseteq \Sigma^{\leq N}K$, i.e., L is not ℓ -stable with respect to K .

Next consider the following family of sublanguages $\{L_i\}_{i \in \mathcal{N}}$ of L with $L_i = a^ib^*$ for each $i \in \mathcal{N}$. Then it is clear that for each $i \in \mathcal{N}$, L_i is controllable (since $\Sigma_c = \Sigma$) and also ℓ -stable (since $L_i \subseteq \Sigma^{\leq i}K$) sublanguage of L . But $\bigcup_{i \in \mathcal{N}} L_i = L$ is not ℓ -stable with respect to K ; thus showing that ℓ -stability is not preserved under union. \square

The implication of Proposition 5.1 is that if the plant behavior L is not ℓ -stable with respect to the desired eventual behavior K , then the minimally restrictive stabilizing supervisor, which will restrict the plant behavior to the supremal ℓ -stable sublanguage of L , cannot in general be constructed. Next we define a weaker notion of language stability that we call *weak ℓ -stability*, which is preserved under union so that the minimally restrictive stabilizing supervisor can be constructed.

DEFINITION 5.2. A language $L \subseteq \Sigma^*$ is said to be *weakly ℓ -stable* with respect to another language $K \subseteq \Sigma^*$ if $L \subseteq \Sigma^*K$. If there exists a nonempty controllable sublanguage $H \subseteq L$ such that H is weakly ℓ -stable with respect to K , then L is said to be *weakly ℓ -stabilizable* with respect to K .

Thus if L is weakly ℓ -stable with respect to K , then every string in L after removing a prefix from it matches some string in K . Note that here no uniform bound on the size of the prefix to be removed from a string in L is assumed.

Remark 5.3. Since $\Sigma^{\leq N} \subseteq \Sigma^*$ for any $N \in \mathcal{N}$, it follows that ℓ -stability implies weak ℓ -stability. However, the converse does not hold in general. Consider, for example, the languages $L = a^*b^*$ and $K = b^*$ defined over the event set $\Sigma = \{a, b\}$. Then as stated in the proof of Proposition 5.1, L is not ℓ -stable with respect to K . But clearly L is weakly ℓ -stable with respect to K , for $a^*b^* \subseteq \Sigma^*b^*$.

The following result, analogous to that stated in Theorem 4.17, holds also for weak ℓ -stabilizability.

THEOREM 5.4. *If L is prefix closed, then L is weakly ℓ -stabilizable with respect to K if and only if $L \cap \Sigma_u^*$ is nonempty and weakly ℓ -stable with respect to K .*

Proof. The proof is similar to the proof of Theorem 4.17. \square

Next we discuss how to verify weak ℓ -stability and weak ℓ -stabilizability of a given

plant behavior with respect to its desired eventual behavior. Let

$$P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m), \quad V \stackrel{\text{def}}{=} (Q, \Sigma, \delta, q_0, Q_m)$$

be the minimal SMs recognizing the languages L, K , respectively. Assuming that the languages L, K , are regular, let m, n be the number of states in P, V , respectively. An SM that recognizes Σ^*K is constructed by first adding the self-loop corresponding to Σ^* at the initial state of V and then converting it to a deterministic SM. Let this SM be denoted by V' ; then the number of states in V' is 2^n .

Remark 5.5. The weak ℓ -stability of L with respect to K can be verified by determining whether $L_m(P) \subseteq L_m(V')$. Since the number of states in P, V' is $m, 2^n$, respectively, the computational complexity of verifying weak ℓ -stability of L with respect to K is $O(m2^n)$. It also follows, in view of Theorem 5.4, that the computational complexity of testing weak ℓ -stabilizability of prefix-closed L with respect to K is again $O(m2^n)$.

Since ℓ -stability (ℓ -stabilizability) implies weak ℓ -stability (weak ℓ -stabilizability), the condition in Proposition 4.19 (Proposition 4.21) is sufficient for weak ℓ -stability (weak ℓ -stabilizability). Thus Algorithm 4.20 (Algorithm 4.22) can be employed to test this sufficient condition for weak ℓ -stability (weak ℓ -stabilizability), the computational complexity of which is polynomial in m, n .

Next we prove that weak ℓ -stability is preserved under union, i.e., the supremal weakly ℓ -stable sublanguage of a given language exists.

PROPOSITION 5.6. *The supremal weakly ℓ -stable sublanguage of a given language exists and is unique.*

Proof. Let L, K denote the plant and desired eventual behavior, respectively. Let Λ be an indexing set such that the family of weakly ℓ -stable sublanguages of L is given by $\{L_\lambda\}_{\lambda \in \Lambda}$, i.e., L_λ is weakly ℓ -stable sublanguage of L for each $\lambda \in \Lambda$. Such a family is nonempty because \emptyset is weakly ℓ -stable sublanguage of L . Consider the language $H \stackrel{\text{def}}{=} \bigcup_{\lambda \in \Lambda} L_\lambda$; then clearly $H \subseteq L$ and H is weakly ℓ -stable. The last assertion follows from the fact that $L_\lambda \subseteq \Sigma^*K$ for each $\lambda \in \Lambda$ which implies that $\bigcup_{\lambda \in \Lambda} L_\lambda = H \subseteq \Sigma^*K$. This completes the proof of Proposition 5.6. \square

COROLLARY 5.7. *The supremal controllable and weakly ℓ -stable sublanguage of a given language exists and is unique.*

Proof. The proof follows from Proposition 5.6 and the fact that controllability is preserved under union [21], [20]. \square

We have proved the existence and uniqueness of the supremal controllable and weakly ℓ -stable sublanguage of a given language. Next we present a closed-form expression for it. We use the notation H^\dagger to denote the supremal controllable sublanguage of a given language $H \subseteq \Sigma^*$ [20,1,10].

THEOREM 5.8. *Let $L, K \subseteq \Sigma^*$ denote the plant and desired eventual behavior, respectively. Then the supremal controllable and weakly ℓ -stable sublanguage of L is given by $(L \cap \Sigma^*K)^\dagger$.*

Proof. Let $H \subseteq \Sigma^*$ denote the supremal controllable and weakly ℓ -stable sublanguage of L with respect to K . Then we must show that $H = (L \cap \Sigma^*K)^\dagger$.

First we show that $(L \cap \Sigma^*K)^\dagger \subseteq H$. Since H is the supremal controllable and weakly ℓ -stable sublanguage of L , it suffices to show that $(L \cap \Sigma^*K)^\dagger$ is a controllable and weakly ℓ -stable sublanguage of L . By its definition, $(L \cap \Sigma^*K)^\dagger$ is a controllable sublanguage of L . Also, since $(L \cap \Sigma^*K)^\dagger \subseteq L \cap \Sigma^*K \subseteq \Sigma^*K$, it follows that $(L \cap \Sigma^*K)^\dagger$ is weakly ℓ -stable with respect to K . Thus $(L \cap \Sigma^*K)^\dagger$ is a controllable and weakly ℓ -stable sublanguage of L .

Next we prove that $H \subseteq (L \cap \Sigma^* K)^\dagger$. Since H is weakly ℓ -stable, it follows that $H \subseteq \Sigma^* K$; also, $H \subseteq L$, hence $H \subseteq L \cap \Sigma^* K$. Note that H is controllable also. Thus H is controllable and is contained in $L \cap \Sigma^* K$. Since $(L \cap \Sigma^* K)^\dagger$ is the supremal controllable sublanguage contained in $L \cap \Sigma^* K$, it follows that $H \subseteq (L \cap \Sigma^* K)^\dagger$. \square

Thus if L is not ℓ -stable with respect to K , but is weakly ℓ -stabilizable with respect to K , then a minimally restrictive stabilizing supervisor can be constructed so that the behavior of the closed-loop system is given by $(L \cap \Sigma^* K)^\dagger$. Note that the result of Theorem 5.8 is not surprising, as we are interested in finding the supremal sublanguage $H \subseteq L$ such that H is weakly stable, i.e., $H \subseteq \Sigma^* K$ and H is controllable. Since $H \subseteq L$ and $H \subseteq \Sigma^* K$, it follows that $H \subseteq L \cap \Sigma^* K$. Thus we are interested in finding the supremal controllable sublanguage $H \subseteq L \cap \Sigma^* K$, which obviously equals $(L \cap \Sigma^* K)^\dagger$. This, however, offers an alternative interpretation of minimally restrictive weakly stabilizing supervisors: the problem of finding the minimally restrictive weakly stabilizing supervisor for a plant with behavior L and *desired eventual behavior* K is equivalent to the problem of finding the minimally restrictive supervisor for the same plant with *desired behavior* $L \cap \Sigma^* K$. Hence techniques developed in [21], [20], [1], [11], etc., can be used to solve the problem.

6. Stability of sequential behavior. Thus far we have discussed the stability of the *finite* behavior of a DEDS. We will show how the notions of ℓ -stability and ℓ -stabilizability defined above can be easily generalized to describe the stability of *infinite* or *sequential* behaviors of DEDSs. In this section, we introduce the notion of ω -stability for formally describing the the notion of eventual sequential behavior.

In [19], [22], [13], [12], [23] the supervisory control problem for controlling the sequential behavior of a DEDS is studied, and conditions under which a supervisor can be constructed so that the sequential behavior of the controlled system is equal to some desired sequential behavior are obtained. As discussed above, such a control problem formulation may lead to synthesis of a very restrictive supervisor. In some cases, it might suffice to design a supervisor that would ensure that the sequential behavior of the controlled system is eventually contained in the desired sequential behavior. So we introduce the notion of the desired eventual sequential behavior and obtain conditions under which the plant's sequential behavior is eventually contained in this sequential behavior. We follow the framework of [19] for addressing the supervisory control problem of sequential behavior.

Let Σ^ω denote the set of all infinite strings of events belonging to Σ . An *infinite* or ω -*language* is a sublanguage of Σ^ω . Let $e^n \in \Sigma^*$ denote the prefix of size n of the infinite string $e \in \Sigma^\omega$. A suitable metric can be defined on the space Σ^ω [7]. Given two infinite strings $e_1, e_2 \in \Sigma^\omega$, the distance $d(e_1, e_2)$ between the two infinite strings is defined to be

$$d(e_1, e_2) \stackrel{\text{def}}{=} \begin{cases} 1/(n+1), & \text{if } e_1^n = e_2^n \text{ and } e_1^{n+1} \neq e_2^{n+1} \ (n \in \mathcal{N}) \\ 0, & \text{if } e_1 = e_2. \end{cases}$$

Given a language $L \subseteq \Sigma^*$, its *limit*, denoted as L^∞ , is the ω -language defined as

$$L^\infty \stackrel{\text{def}}{=} \{e \in \Sigma^\omega \mid e^n \in L \text{ for infinitely many } n \in \mathcal{N}\}.$$

We will use $t \leq s$ to denote that $t \in \Sigma^*$ is a prefix of $s \in \Sigma^* \cup \Sigma^\omega$. If t is a proper prefix of s , then it is written as $t < s$. Given an infinite sequence of strings $s_1 < s_2 < \dots < s_n < \dots$ with $s_n \in \Sigma^*$ for each n , there exists a unique infinite string $e \in \Sigma^\omega$ such that $s_n < e$ for each n . In this case, the infinite string e is also written

as $e = \lim_{n \rightarrow \infty} s_n$. Given an ω -language $\mathcal{L} \subseteq \Sigma^\omega$, its *prefix*, denoted by $pr\mathcal{L}$, is the following language:

$$pr\mathcal{L} \stackrel{\text{def}}{=} \{s \in \Sigma^* \mid \exists e \in \mathcal{L} \text{ s.t. } s < e\}.$$

Note that $pr\mathcal{L} = pr\bar{\mathcal{L}}$, where $\bar{\mathcal{L}}$ denotes the topological closure³ of \mathcal{L} in the metric space (Σ^ω, d) [7]. It can be proved [7] that for a ω -language $\mathcal{L} \subseteq \Sigma^\omega$,

$$(pr\mathcal{L})^\infty = \bar{\mathcal{L}}.$$

With the above preliminary notions we can address the issue of stability of the infinite behavior of a given DEDS. Let $P \equiv (X, \Sigma, \alpha, x_0, X_m)$ denote the plant. Then as defined above, $L_m(P), L(P) \subseteq \Sigma^*$ denote its (finite) marked, generated languages respectively. The ω -language generated by P , denoted by $\mathcal{L}(P)$, is defined to be

$$\mathcal{L}(P) \stackrel{\text{def}}{=} \{e \in (L(P))^\infty \mid \exists \text{ infinitely many } n \in \mathcal{N} \text{ s.t. } \alpha(e^n, x_0) \in X_m\} = (L_m(P))^\infty.$$

Note that the ω -language $\mathcal{L}(P)$ generated by P as defined above is also the ω -language generated by P viewed as a Büchi automaton [7]. P is said to be *nonblocking* if $pr\mathcal{L}(P) = L(P)$. Let $S \equiv (Y, \Sigma, \beta, y_0, Y_m)$ denote the supervisor that controls P by synchronization as defined above. Then the ω -language generated by the closed-loop system $P \square S$ is defined to be

$$\mathcal{L}(P \square S) \stackrel{\text{def}}{=} (L(P \square S))^\infty \cap \mathcal{L}(P).$$

Let $\mathcal{K} \subseteq \mathcal{L}(P)$ be the desired ω -language. It is shown in [19] that a complete, nonblocking supervisor exists for achieving the desired sequential behavior if and only if \mathcal{K} is ω -controllable with respect to P .

DEFINITION 6.1. An ω -language $\mathcal{K} \subseteq \Sigma^\omega$ is said to be ω -controllable with respect to the plant P if $pr\mathcal{K}$ is controllable with respect to P , and \mathcal{K} is topologically closed with respect to $\mathcal{L}(P)$; i.e.,

1. $pr(\mathcal{K})\Sigma_u \cap L(P) \subseteq pr\mathcal{K}$, and
2. $\bar{\mathcal{K}} \cap \mathcal{L}(P) = \mathcal{K}$.

It is further shown in [19] that if \mathcal{K} is not ω -controllable, but is topologically closed with respect to $\mathcal{L}(P)$, then the *supremal ω -controllable* sublanguage, denoted by \mathcal{K}^\uparrow , of \mathcal{K} exists.⁴ Thus the construction of the *minimally restrictive supervisor* is possible. A closed-form expression for the supremal ω -controllable sublanguage, as well as an efficient algorithm for computing it, is presented in [13], [12].

Next, let $\mathcal{K} \subseteq \Sigma^\omega$ represent the desired eventual sequential behavior of the plant $P \equiv (X, \Sigma, \alpha, x_0, X_m)$. The notion of ω -stability is defined as follows.

DEFINITION 6.2. The plant sequential behavior $\mathcal{L}(P)$ is said to be ω -stable with respect to the desired eventual sequential behavior \mathcal{K} if there exists an integer $N \in \mathcal{N}$ such that $\mathcal{L}(P) \subseteq \Sigma^{\leq N}\mathcal{K}$. $\mathcal{L}(P)$ is said to be ω -stabilizable with respect to \mathcal{K} if there exists a nonempty ω -controllable sublanguage $\mathcal{H} \subseteq \mathcal{L}(P)$ such that \mathcal{H} is ω -stable with respect to \mathcal{K} .

³ The notation $\bar{\mathcal{L}}$ is used to denote topological closure whenever $\mathcal{L} \subseteq \Sigma^\omega$, and the notation \bar{L} is used to denote the prefix closure whenever $L \subseteq \Sigma^*$.

⁴ The notation \mathcal{K}^\uparrow is used to denote the supremal ω -controllable sublanguage of $\mathcal{K} \subseteq \Sigma^\omega$, and the notation K^\uparrow is used to denote the supremal controllable sublanguage of $K \subseteq \Sigma^*$.

Let $e \in \Sigma^\omega$ be an infinite string and for each $n \in \mathcal{N}$, let $f_n \in \Sigma^\omega$ be such that $e = e^n f_n$. Then the *projection* operator $\Pi_n : \Sigma^\omega \rightarrow \Sigma^\omega$ ($n \in \mathcal{N}$) is defined in the following manner:

$$\Pi_n(e) = f_n.$$

In other words, given an infinite string $e \in \Sigma^\omega$, its projection $\Pi_n(e)$ is obtained by deleting its prefix of size n from it. Thus if $\mathcal{L}(P)$ is ω -stable with respect to \mathcal{K} , then for each $e \in \mathcal{L}(P)$ there exists an integer $n_e \leq N$ such that $\Pi_{n_e}(e) \in \mathcal{K}$. In other words, each infinite string in $\mathcal{L}(P)$ after removing a prefix of size at most N matches an infinite string in \mathcal{K} . The ω -language \mathcal{K} thus can be thought of to be representing the desired eventual sequential behavior. If $\mathcal{L}(P)$ is not ω -stable but ω -stabilizable with respect to \mathcal{K} , then there exists a nonempty ω -controllable sublanguage $\mathcal{H} \subseteq \mathcal{L}(P)$ that is ω -stable with respect to \mathcal{K} also. Thus a nonblocking and complete [19] supervisor, which can restrict the sequential behavior of the plant to \mathcal{H} that “stabilizes” to the desired eventual sequential behavior \mathcal{K} , can be constructed.

6.1. Tests for ω -stability and ω -stabilizability. In this subsection we show that under certain assumptions ω -stability can be tested by performing the test for ℓ -stability. First, we define the notion of *complete* languages, which is useful in the context of studying the stability of infinite behaviors.

DEFINITION 6.3. Consider a language $L \subseteq \Sigma^*$. A string $s \in L$ is said to have an *extension* in L if there exists a $t \in L$ such that $s < t$. L is said to be *complete*⁵ if for every string $s \in L$, there exists an extension in L .

Note that a language is complete if and only if a trim SM recognizing it is *live* (has at least one transition defined at each of its states) [13]. First, we show that ℓ -stability of a given language with respect to another implies ω -stability of the limit of the given language with respect to the limit of the other.

THEOREM 6.4. Consider $L, K \subseteq \Sigma^*$. If L is ℓ -stable with respect to K , then L^∞ is ω -stable with respect to K^∞ .

We prove the following lemma before proving the result of Theorem 6.4.

LEMMA 6.5. Consider $L \subseteq \Sigma^*$. Then for any $N \in \mathcal{N}$, $(\Sigma^{\leq N} L)^\infty = \Sigma^{\leq N} L^\infty$.

Proof. First, we show that $\Sigma^{\leq N} L^\infty \subseteq (\Sigma^{\leq N} L)^\infty$. Pick $e \in \Sigma^{\leq N} L^\infty$. Then e can be written as $e = e^n f$, where $n \leq N$ and $f \in L^\infty$. Thus there exist infinitely many $m \in \mathcal{N}$ such that $f^m \in L$. Then the strings $e^n f^m \in \Sigma^{\leq N} L$ for each $m \in \mathcal{N}$. Hence $\lim_{m \rightarrow \infty} e^n f^m \in (\Sigma^{\leq N} L)^\infty$. Also, since $e^n f^1 < e^n f^2 < \dots < e^n f^m < \dots < e$, it follows that $\lim_{m \rightarrow \infty} e^n f^m = e$, which shows that $e \in (\Sigma^{\leq N} L)^\infty$.

Next, we show that $(\Sigma^{\leq N} L)^\infty \subseteq \Sigma^{\leq N} L^\infty$. Pick $e \in (\Sigma^{\leq N} L)^\infty$. Then there exist infinitely many $n \in \mathcal{N}$ such that $e^n \in \Sigma^{\leq N} L$. Thus each e^n can be written as $e^n = u_n v_n$, where $u_n \in \Sigma^{\leq N}$ and $v_n \in L$. Since the set $\Sigma^{\leq N}$ is finite, it follows that there exists at least one integer $n_0 \in \mathcal{N}$ such that $u_{n_0} = u_n$ for infinitely many n . Let $\{n_k\}_{k \in \mathcal{N}}$ be a subsequence such that $u_{n_1} = u_{n_2} = \dots = u_{n_k} = \dots = u_{n_0}$. Then $e^{n_k} = u_{n_0} v_{n_k}$ for each $k \in \mathcal{N}$. Hence $e = \lim_{k \rightarrow \infty} e^{n_k} = u_{n_0} \lim_{k \rightarrow \infty} v_{n_k}$. Since $u_{n_0} \in \Sigma^{\leq N}$ and $v_{n_k} \in L$ for each $k \in \mathcal{N}$, it follows that $e \in \Sigma^{\leq N} L^\infty$. \square

Proof of Theorem 6.4. Since L is ℓ -stable with respect to K , there exists an integer $N \in \mathcal{N}$ such that $L \subseteq \Sigma^{\leq N} K$. Hence, by taking limits on both sides of the last inclusion, we obtain $L^\infty \subseteq (\Sigma^{\leq N} K)^\infty$. It then follows from Lemma 6.5 that $L^\infty \subseteq \Sigma^{\leq N} K^\infty$, which shows that L^∞ is ω -stable with respect to K^∞ . \square

⁵ Completeness is also defined to be a property of supervisors; here we define it to be a property of languages. The two definitions are unrelated and not to be confused with.

Example 6.6. Consider languages L, K_1 of Example 4.5. Then $L^\infty = ((ab)^*cd^*)^\omega = (ab)^*cd^\omega$ and $(K_1)^\infty = (d^* + bab(ab)^*cd^*)^\infty = d^\omega + bab(ab)^*cd^\omega$. Using arguments similar to those in Example 4.5, it can be easily verified that $L^\infty \subseteq \Sigma^{\leq 3}(K_1)^\infty$. This shows, as expected from the result of Theorem 6.4, that L^∞ is ω -stable with respect to $(K_1)^\infty$. However, the converse of Theorem 6.4 does not hold in general. Consider, for example, languages $L, K \subseteq \{a, b\}^*$: $L = (ab)^*$ and $K = (ba)^*$. Then $L^\infty = (ab)^\omega$ and $K^\infty = (ba)^\omega$. Since $(ab)^\omega = a(ba)^\omega$, it is obvious that L^∞ is ω -stable with respect to K^∞ ($L^\infty = aK^\infty$). It can also be easily checked that L is not ℓ -stable with respect to K : a string in L ends with the symbol b , whereas a string in K ends with the symbol a . Thus, given any string in L , no suffix of it matches any string in K .

Next we prove that under certain assumptions the converse of Theorem 6.4 holds.

THEOREM 6.7. *Consider $L, K \subseteq \Sigma^*$. Assume that L is complete and K is prefix closed. Then ω -stability of L^∞ with respect to K^∞ implies ℓ -stability of L with respect to K .*

Before proving the result of Theorem 6.7, we prove the following lemma.

LEMMA 6.8. *Consider two languages $L_1, L_2 \subseteq \Sigma^*$. Assume that L_1 is complete and L_2 is closed. Then $(L_1)^\infty \subseteq (L_2)^\infty$ if and only if $L_1 \subseteq L_2$.*

Proof. It is clear that $L_1 \subseteq L_2$ implies $L_1^\infty \subseteq L_2^\infty$. Hence it suffices to show that if $(L_1)^\infty \subseteq (L_2)^\infty$, then $L_1 \subseteq L_2$. Pick $s \in L_1$. Since L_1 is complete, there exists a sequence of strings $s_1 < s_2 < \dots < s_n < \dots$ such that $s_n \in L_1$ for each $n \in \mathcal{N}$ and $s < s_1$. Let $e = \lim_{n \rightarrow \infty} s_n$; then $e \in (L_1)^\infty$. It then follows from the assumption that $e \in (L_2)^\infty$. Hence there exist infinitely many $n \in \mathcal{N}$ such that $e^n \in L_2$. Pick $m \in \mathcal{N}$ such that $s < e^m$. Since $e^m \in L_2$ and L_2 is closed, it follows that $s \in L_2$. \square

Proof of Theorem 6.7. Assume that L^∞ is ω -stable with respect to K^∞ . Then there exists an integer $N \in \mathcal{N}$ such that $L^\infty \subseteq \Sigma^{\leq N}K^\infty$. Thus it follows from Lemma 6.5 that $L^\infty \subseteq (\Sigma^{\leq N}K)^\infty$. Note that since $\Sigma^{\leq N}$ is closed, and prefix closure is preserved under concatenation of languages, $\Sigma^{\leq N}K$ is a closed language (by assumption K is closed). Since L is complete (by assumption) and $\Sigma^{\leq N}K$ is closed, we obtain from Lemma 6.8 that $L^\infty \subseteq (\Sigma^{\leq N}K)^\infty$ if and only if $L \subseteq \Sigma^{\leq N}K$. \square

Example 6.9. Consider the languages $L = (ab)^*$ and $K = (ba)^*$ of Example 6.6. It was noted in Example 6.6 that L^∞ is ω -stable with respect to K^∞ ; however, L is not ℓ -stable with respect to K . The reason is that although L is a complete language, K is not prefix closed. Let us replace K by its prefix closure, i.e., consider $K' = \overline{K} = (ba)^* = (ba)^* + b(ab)^*$. Then clearly L is ℓ -stable with respect to K' ($L = (ab)^* = ab + ab(ab)^* \subseteq \Sigma^{\leq 2}K'$).

The results of Theorem 6.4 and Theorem 6.7 can be combined to arrive at a test for ω -stability based on the test for ℓ -stability (Theorem 4.12).

THEOREM 6.10. *Let $\mathcal{L}(P) = (L_m(P))^\infty \subseteq \Sigma^\omega$ denote the plant ω -behavior and $\mathcal{K} \subseteq \Sigma^\omega$ denote the desired eventual behavior. If P is live and \mathcal{K} is topologically closed, then $\mathcal{L}(P)$ is ω -stable with respect to \mathcal{K} if and only if $L_m(P)$ is ℓ -stable with respect to $pr\mathcal{K}$.*

Proof. Since P is live, $L_m(P)$ is complete. Also, since \mathcal{K} is topologically closed, $\mathcal{K} = \overline{\mathcal{K}} = (pr\mathcal{K})^\infty$. Thus $\mathcal{L}(P)$ is the limit of the complete language $L_m(P)$ and \mathcal{K} is the limit of the prefix-closed language $pr\mathcal{K}$. Hence it follows from Theorem 6.4 and Theorem 6.7 that $\mathcal{L}(P)$ is ω -stable with respect to \mathcal{K} if and only if $L_m(P)$ is ℓ -stable with respect to $pr\mathcal{K}$. \square

Next we relate the notion of ω -stabilizability to that of ω -stability through the following theorem.

THEOREM 6.11. $\mathcal{L}(P)$ is ω -stabilizable with respect to \mathcal{K} if and only if $\mathcal{L}(P) \cap \Sigma_u^\omega$ is nonempty and ω -stable with respect to \mathcal{K} , where $\Sigma_u^\omega = (\Sigma_u^*)^\omega$.

Proof. We first show that $\mathcal{L}(P) \cap \Sigma_u^\omega$ is the infimal ω -controllable sublanguage of $\mathcal{L}(P)$, i.e., it is the sequential behavior of P under the control of maximally restrictive complete and nonblocking supervisor [19]. Consider the supervisor that disables all the controllable events in P . Then the behavior of the closed-loop system under this control law is given by $L(P) \cap \Sigma_u^*$. Hence the sequential behavior of the closed-loop system is given by $(L(P) \cap \Sigma_u^*)^\omega \cap \mathcal{L}(P) = (L(P))^\omega \cap (\Sigma_u^*)^\omega \cap \mathcal{L}(P) = \mathcal{L}(P) \cap \Sigma_u^\omega$, where the first equality follows from the fact that $L(P), \Sigma_u^*$ are both closed languages and the second equality follows from the fact that $\mathcal{L}(P) \subseteq (L(P))^\omega$ and $(\Sigma_u^*)^\omega = \Sigma_u^\omega$. Note that the supervisor that disables all the controllable transitions in P is complete (it never disables any uncontrollable transition) and nonblocking (since $pr(\mathcal{L}(P) \cap \Sigma_u^\omega) = L(P) \cap \Sigma_u^*$). Hence $\mathcal{L}(P) \cap \Sigma_u^\omega$ is ω -controllable [19]. Since it is the sequential behavior under the maximally restrictive complete and nonblocking control law, if $\mathcal{H} \subseteq \mathcal{L}(P)$ is any ω -controllable sublanguage of $\mathcal{L}(P)$, then $\mathcal{L}(P) \cap \Sigma_u^\omega \subseteq \mathcal{H}$.

Assume then that $\mathcal{L}(P)$ is ω -stabilizable with respect to \mathcal{K} . Then by the definition of ω -stabilizability, there exists a nonempty ω -controllable sublanguage $\mathcal{H} \subseteq \mathcal{L}(P)$ and an integer $N \in \mathcal{N}$ such that $\mathcal{H} \subseteq \Sigma^{\leq N} \mathcal{K}$. Since $\mathcal{L}(P) \cap \Sigma_u^\omega \subseteq \mathcal{H}$, it follows that $\mathcal{L}(P) \cap \Sigma_u^\omega \subseteq \Sigma^{\leq N} \mathcal{K}$; which shows that $\mathcal{L}(P) \cap \Sigma_u^\omega$ is ω -stable with respect to \mathcal{K} .

Assume next that $\mathcal{L}(P) \cap \Sigma_u^\omega$ is nonempty and ω -stable with respect to \mathcal{K} . Since $\mathcal{L}(P) \cap \Sigma_u^\omega \subseteq \mathcal{L}(P)$ and is ω -controllable (proved above), it follows that $\mathcal{L}(P)$ is ω -stabilizable with respect to \mathcal{K} . \square

Remark 6.12. A necessary condition for ω -stability is obtained using an equivalence relation on the space Σ^ω introduced in Appendix B. It is also shown in Appendix B that if a weaker definition of ω -stability is used the necessary condition obtained in terms of the equivalence relation is also a sufficient condition.

7. Conclusion. In this paper, we have introduced the notions of stability and stabilizability of DEDSs in terms of their behavior. In many situations, since the behavior rather than the states of the system is observed directly, it is more natural to study the stability of systems in terms of their behavior. Also, in some cases, it might be desired that the eventual (rather than the whole) behavior of the system be legal, so it is necessary to define formally the notion of language stability. Earlier works concerning stability of DEDSs [18], [2], [3] are all based in terms of the states of the systems and can be viewed as a special case of the work presented here (refer to Proposition 4.4). The earlier works [18], [2], [3] on stability in terms of states assume the control to be of static feedback type; however, more general supervisors that exercise dynamic feedback have been used here for making the systems ℓ -stable.

We have shown that the problem of determining ℓ -stability (ℓ -stabilizability) of a given language with respect to another language is equivalent to another problem posed in terms of the reversal of languages (refer to Corollary 4.10) and have provided a solution to this equivalent problem (refer to Theorems 4.12 and 4.17). We have also provided an upper bound to the value of the integer N in the definition of ℓ -stability (ℓ -stabilizability) using the solution to the equivalent problem (refer to Corollary 4.10). Next we have presented a weaker notion of language stability in which no uniform upper bound on the length of the prefix to be removed from a string in a language (for it to be ℓ -stable with respect to another language) exists and have provided the construction of the *minimally restrictive supervisor* [10], [21], [20], [11] to ℓ -stabilize a given language in this weaker sense of language stability.

The notion of ℓ -stability and ℓ -stabilizability is then generalized to describe the

notion of stability of sequential behavior of DEDSs and the notions of ω -stability and ω -stabilizability are introduced in this context. We have introduced an equivalence relation on the space of infinite strings and have obtained a necessary condition of ω -stability in terms of this relation. A necessary and sufficient condition for ω -stability is obtained in terms of ℓ -stability, which is used to arrive at tests for ω -stability and ω -stabilizability.

A. Algorithm for constructing $\Omega(\hat{X})$ and $\Lambda(\hat{X})$. As before, let

$$P \stackrel{\text{def}}{=} (X, \Sigma, \alpha, x_0, X_m)$$

be the plant and $\hat{X} \subseteq X$ be the set of legal states. The following algorithm can be used to compute $\Omega(\hat{X})$ (we assume that the plant P has finite number of states so that the algorithm terminates in finite number of steps).

ALGORITHM A.1.

1. Initiation step:

Set $\Omega_{-1}(\hat{X}) = \emptyset, \Omega_0(\hat{X}) = \hat{X}$, and $k = 0$.

2. Iteration step:

- (a) Let $X_k \subseteq X$ be the set of states from which $\Omega_k(\hat{X}) - \Omega_{k-1}(\hat{X})$ can be reached in a single transition, i.e.,

$$X_k = \{x \in X \mid \exists \sigma \in \Sigma \text{ s.t. } \alpha(\sigma, x) \in \Omega_k(\hat{X}) - \Omega_{k-1}(\hat{X})\}.$$

Determine the set X_k by considering the SM $P^{-1} \stackrel{\text{def}}{=} (X, \Sigma, \alpha^{-1}, x_0, X_m)$, where $\alpha^{-1}(\sigma, x_2) \stackrel{\text{def}}{=} \{x_1 \in X \mid \alpha(\sigma, x_1) = x_2\}$ (P^{-1} is the SM obtained by reversing all the transitions of P), and by finding the states that can be reached from $\Omega_k(\hat{X}) - \Omega_{k-1}(\hat{X})$ by a single transition in P^{-1} .

- (b) Consider $x \in X_k$. If all the transitions from x lead to $\Omega_k(\hat{X})$, then $\Omega_{k+1}(\hat{X}) = \Omega_k(\hat{X}) \cup \{x\}$. Repeat this for all $x \in X_k$. Thus, if all the transitions from a state $x \in X_k$ lead to states in $\Omega_k(\hat{X})$, then x is a strongly attractable state, i.e.,

$$\Omega_{k+1}(\hat{X}) = \Omega_k(\hat{X}) \cup \{x \in X_k \mid \alpha(\sigma, x) \in \Omega_k(\hat{X}) \text{ for all } \sigma \in \Sigma(P)(x)\},$$

where $\Sigma(P)(x) \subseteq \Sigma$ is the set of all the transitions that are defined in the state $x \in X$ in P and is given by $\Sigma(P)(x) = \{\sigma \in \Sigma \mid \alpha(\sigma, x)!\}$.

3. Termination step:

If $\Omega_{k+1}(\hat{X}) = \Omega_k(\hat{X})$, then stop and set $\Omega(\hat{X}) = \Omega_k(\hat{X})$; else set $k = k + 1$ and go to step 2.

THEOREM A.2. *Algorithm A.1 computes the region of strong attraction $\Omega(\hat{X})$ of the set of legal states $\hat{X} \subseteq X$.*

Proof. The proof that Algorithm A.1 computes $\Omega(\hat{X})$ is based on the following two facts:

First, the above algorithm computes $\Omega(\hat{X})$ if in step 2, $\Omega_k(\hat{X}) - \Omega_{k-1}(\hat{X})$ is replaced by $\Omega_k(\hat{X})$ (for proof refer to [18, Prop. 2.7]).

Second, at the end of the k th iteration, to determine the states that might be strongly attractable, we just need to consider the states that have transitions leading into the set $\Omega_{k+1}(\hat{X}) - \Omega_k(\hat{X})$ (rather than into the set $\Omega_{k+1}(\hat{X})$) in P , so that the replacement as described above is justified (see Fig. 6). In other words, we must show that at the end of k th iteration, if all the transitions in $\Sigma(P)(x)$ from the state

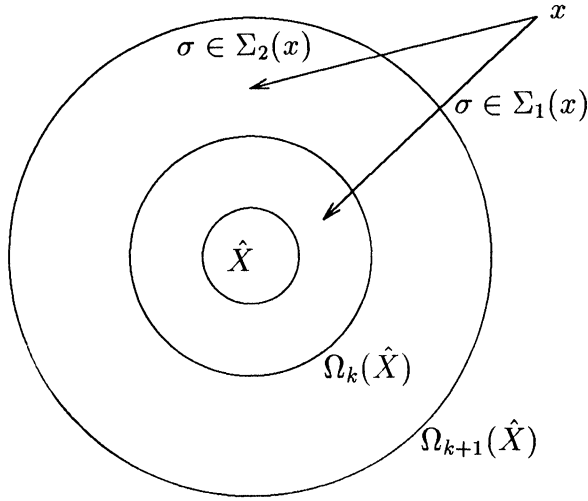


FIG. 6. Constructing region of strong attraction.

$x \in X - \Omega_{k+1}(\hat{X})$ lead to the set $\Omega_{k+1}(\hat{X})$, then there exists $\sigma \in \Sigma(x)$ such that $\alpha(\sigma, x) \in \Omega_{k+1}(\hat{X}) - \Omega_k(\hat{X})$. To show this, we first partition $\Sigma(P)(x)$ into the set $\Sigma_1(P)(x) \cup \Sigma_2(P)(x)$, the set $\Sigma_1(P)(x)$ of transitions leading to $\Omega_k(\hat{X})$ and the set $\Sigma_2(P)(x)$ of transitions leading to $\Omega_{k+1}(\hat{X}) - \Omega_k(\hat{X})$. Then it is enough to show that the set $\Sigma_2(x)$ is nonempty. Assume that it is empty; then $x \in \Omega(\Omega_k(\hat{X}))$; therefore it belongs to the set $\Omega_{k+1}(\hat{X})$, which is contradictory to the fact that $x \in X - \Omega_{k+1}(\hat{X})$. This proves the second claim. \square

Remark A.3. To determine the region of weak attraction $\Lambda(\hat{X})$ of \hat{X} , we replace step 2(b) in the iteration step of the previous algorithm by the following step 2(b'):

2(b') Consider $x \in X_k$. If all the uncontrollable transitions from x lead to $\Omega_k(\hat{X})$, then $\Omega_{k+1}(\hat{X}) = \Omega_k(\hat{X}) \cup \{x\}$, i.e.,

$$\Omega_{k+1}(\hat{X}) = \Omega_k(\hat{X}) \cup \{x \in X_k \mid \alpha(\sigma, x) \in \Omega_k(\hat{X}) \text{ for all } \sigma \in \Sigma_u(P)(x)\},$$

where $\Sigma_u(P)(x) = \Sigma(P)(x) \cap \Sigma_u$.

This can be tested by considering the transitions in $P|_{\Sigma_u}$ (P with all its controllable transitions deleted). Formally, $P|_{\Sigma_u} \stackrel{\text{def}}{=} (X, \Sigma_u, \alpha|_{\Sigma_u \times X}, x_0, X_m)$.

This would result in the construction of the region of weak attraction $\Lambda(\hat{X})$ of \hat{X} . Note that with an abuse of notation we have used $\Omega_k(\hat{X})$ in the algorithm for determining $\Lambda_k(\hat{X})$.

THEOREM A.4. The time complexity of Algorithm A.1 for constructing $\Omega(\hat{X})$ and $\Lambda(\hat{X})$ is $O(|\Sigma|n)$, where $|\Sigma|$ denotes the number of events in the event set Σ and n is the number of states in P .

Proof. Assume that at the end of k th iteration, the number of transitions (of length one) leading into the set $\Omega_{k+1}(\hat{X}) - \Omega_k(\hat{X})$ from $X - \Omega_{k+1}(\hat{X})$ is e_k . We show that step 2 of the algorithm can be computed in $O(e_k)$ time, as follows.

First, the states in the set X_k can be computed in $O(e_k)$ time, for to determine the states reachable from the states in the set $\Omega_{k+1}(\hat{X}) - \Omega_k(\hat{X})$ by a single transition in P^{-1} , we need consider only the e_k transitions. Second, since there could be at most

e_k such states, the states in the set $\Omega_{k+1}(\hat{X})$ can also be computed in $O(e_k)$ time. This is true because to test whether a state $x \in X_k$ belongs to $\Omega_{k+1}(\hat{X})$ requires only $O(|\Sigma|)$ time, which is constant.

Since the sets $\Omega_{k+1}(\hat{X}) - \Omega_k(\hat{X})$ for each value of k are all disjoint, the transitions (of length one) leading into them from $X - \Omega_{k+1}(\hat{X})$ are also all disjoint. Hence the computational complexity of Algorithm A.1 is of order $O(\sum_k e_k) = O(e)$, where e is the number of transitions in P . Since P is deterministic, $e \leq |\Sigma|n$; hence the theorem follows. Similarly, the complexity of the algorithm for determining $\Lambda(\hat{X})$ is also $O(|\Sigma|n)$. \square

This is significant improvement over the computational complexity of the algorithm given in [2], [3], which is $O(n^2)$. Note that our algorithm requires the construction of the SM P^{-1} , which could be nondeterministic, but has same number of transitions as P .

The above algorithm can also be used to construct the *prestable* and *prestablizable* states of a given *invariant* state set as defined in [18]. In fact, the set of prestable states and the set of prestabilizable states with respect to a given invariant or legal set of states is the same as $\Omega(\hat{X})$ and $\Lambda(\hat{X})$, respectively, where \hat{X} denotes the set of invariant states. The computational complexity of the algorithms provided in [18] is also quadratic in the number of states of P .

B. An equivalence relation on Σ^ω and ω -stability. A necessary condition for ω -stability of a given ω -language with respect to another can be obtained in terms of an equivalence relation defined on the space Σ^ω . In this appendix we define this relation and show its close relation to the notion of ω -stability.

DEFINITION B.1. For $e_1, e_2 \in \Sigma^\omega$, $e_1 \cong e_2$ if and only if there exist $m, n \in \mathcal{N}$ such that $\Pi_m(e_1) = \Pi_n(e_2)$. Note that for each $n \in \mathcal{N}$, $\Pi_n : \Sigma^\omega \rightarrow \Sigma^\omega$ is the map such that for $e \in \Sigma^\omega$, $\Pi_n(e)$ is the infinite string obtained by removing the prefix of length n from e .

THEOREM B.2. *The relation \cong as defined is an equivalence relation.*

Proof. We must show that the relation \cong is reflexive, symmetric, and transitive.

It is clear that for any vector $e_1 \in \Sigma^\omega$, $e_1 \cong e_1$, i.e., \cong is reflexive. Also, if $e_1 \cong e_2$, then clearly $e_2 \cong e_1$ for any two vectors $e_1, e_2 \in \Sigma^\omega$, i.e., \cong is symmetric. It remains to show that the relation \cong is transitive. Pick any $e_1, e_2, e_3 \in \Sigma^\omega$. We will show that $e_1 \cong e_2$ and $e_2 \cong e_3$ implies $e_1 \cong e_3$. Let $m, n, p, q \in \mathcal{N}$ be such that $\Pi_m(e_1) = \Pi_n(e_2)$ and $\Pi_p(e_2) = \Pi_q(e_3)$. We may have either $n \leq p$ or $p \leq n$. If $n \leq p$, then $\Pi_{m+(p-n)}(e_1) = \Pi_q(e_3)$, i.e., $e_1 \cong e_3$; if $p \leq n$, then $\Pi_m(e_1) = \Pi_{q+(n-p)}(e_3)$, i.e., $e_1 \cong e_3$. \square

A necessary condition for ω -stability can be obtained using the equivalence relation defined above.

PROPOSITION B.3. *If plant sequential behavior $\mathcal{L}(P)$ is ω -stable with respect to the desired eventual behavior \mathcal{K} , then for each $e \in \mathcal{L}(P)$, there exists $e' \in \mathcal{K}$ such that $e \cong e'$.*

Proof. Assume $\mathcal{L}(P)$ is ω -stable with respect to \mathcal{K} , i.e., there exist $N \in \mathcal{N}$ such that $\mathcal{L}(P) \subseteq \Sigma^{\leq N}\mathcal{K}$. Then given $e \in \mathcal{L}(P)$, there exists $n \leq N$ and $e' \in \mathcal{K}$ such that $e = e^n e'$. Thus $\Pi_n(e) = e'$, i.e. $e \cong e'$. \square

Remark B.4. Proposition B.3 gives a necessary condition for ω -stability. This condition will be a necessary as well as sufficient condition if a weaker definition of ω -stability is used. Let the projection operator be extended to the space 2^{Σ^ω} in the obvious manner, i.e., for any $n \in \mathcal{N}$, $\Pi_n : 2^{\Sigma^\omega} \rightarrow 2^{\Sigma^\omega}$ is defined to be

$$\Pi_n(\mathcal{L}) = \{e \in \Sigma^\omega \mid \exists e' \in \mathcal{L} \text{ s.t. } \Pi_n(e') = e\},$$

where $\mathcal{L} \subseteq \Sigma^\omega$. We use $\Pi_\star(\cdot)$ to denote the operator $\bigcup_{n \in \mathcal{N}} \Pi_n(\cdot)$. The plant sequential behavior $\mathcal{L}(P)$ is said to be *weakly ω -stable* with respect to the desired eventual sequential behavior \mathcal{K} if $\mathcal{L}(P) \subseteq \Sigma^\star \Pi_\star(\mathcal{K})$. Thus if $\mathcal{L}(P)$ is weakly ω -stable with respect to \mathcal{K} , then for every $e \in \mathcal{L}(P)$ there exist $n, m \in \mathcal{N}$ and $e' \in \mathcal{K}$ such that $\Pi_n(e) = \Pi_m(e')$. It is clear that ω -stability implies weak ω -stability. It can easily be verified that $\mathcal{L}(P)$ is weakly ω -stable with respect to \mathcal{K} if and only if given any $e \in \mathcal{L}(P)$ there exists $e' \in \mathcal{K}$ such that $e \cong e'$.

REFERENCES

- [1] R. D. BRANDT, V. K. GARG, R. KUMAR, F. LIN, S. I. MARCUS, AND W. M. WONHAM, *Formulas for calculating supremal and normal sublanguages*, Systems Control Lett., 15 (1990), pp. 111–117.
- [2] Y. BRAVE AND M. HEYMANN, *On stabilization of discrete event processes*, In IEEE Proceedings of 28th Conference on Decision and Control, Tampa, FL, December, 1989, pp. 2737–2742.
- [3] ———, *On stabilization of discrete event processes*, Tech. Report, Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel, 1989.
- [4] ———, *On optimal attraction in discrete event processes*, Tech. Report CIS Report 9019, Department of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel, 1990.
- [5] H. CHO AND S. I. MARCUS, *On supremal languages of class of sublanguages that arise in supervisor synthesis problems with partial observations*, Math. Control Signals Systems, 2 (1989), pp. 47–69.
- [6] ———, *Supremal and maximal sublanguages arising in supervisor synthesis problems with partial observations*, Math. Systems Theory, 22 (1989), pp. 177–211.
- [7] S. EILENBERG, *Automata, Languages, and Machines: Volume A*, Academic Press, New York, 1974.
- [8] J. E. HOPCROFT AND J. D. ULLMAN, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading, MA, 1979.
- [9] R. KUMAR, *Supervisory Synthesis Techniques for Discrete Event Dynamical Systems: Transition Model Based Approach*, Ph.D. Thesis, Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, 1991.
- [10] R. KUMAR, V. K. GARG, AND S. I. MARCUS, *Supervisory control of discrete event systems: supremal controllable and observable languages*, in Proceedings of 1989 Allerton Conference, Allerton, IL, September, 1989, pp. 501–510.
- [11] ———, *On controllability and normality of discrete event dynamical systems*, Systems Control Lett., 17 (1991), pp. 157–168.
- [12] ———, *On ω -controllability and ω -normality of dedfs*, in Proceedings of 1991 ACC, Boston, MA, June, 1991, pp. 2905–2910.
- [13] ———, *On supervisory control of sequential behaviors*, IEEE Trans. Automat. Control, 37 (1992), pp. 1978–1985.
- [14] S. LAFORTUNE AND E. CHEN, *On the infimal closed and controllable superlanguage of a given language*, IEEE Trans. Automat. Control, 35 (1990), pp. 398–404.
- [15] S. LAFORTUNE AND F. LIN, *On tolerable and desirable behaviors in supervisory control of discrete event systems*, Discrete Event Dynam. System: Theory Appl., 1 (1991), pp. 61–92.
- [16] C. M. OZVEREN, A. S. WILLSKY, AND P. J. ANTSAKLIS, *Output stabilizability of discrete event dynamical systems*, in IEEE Proceedings of 28th Conference on Decision and Control, Tampa, FL, December, 1989, pp. 2719–2724.
- [17] C. M. OZVEREN AND A. S. WILLSKY, *Tracking and restrictability in discrete event dynamical systems*, SIAM J. Control Optim., 30 (1992), pp. 1423–1446.
- [18] C. M. OZVEREN, A. S. WILLSKY, AND P. J. ANTSAKLIS, *Stability and stabilizability of discrete event dynamical systems*, Journal of ACM, 38 (1991), pp. 730–752.
- [19] P. J. RAMADGE, *Some tractable supervisory control problems for discrete event systems modeled by buchi automata*, IEEE Trans. Automat. Control, 34 (1989), pp. 10–19.
- [20] P. J. RAMADGE AND W. M. WONHAM, *On the supremal controllable sublanguage of a given language*, SIAM J. Control Optim., 25 (1987), pp. 637–659.
- [21] ———, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.
- [22] J. G. THISTLE AND W. M. WONHAM, *On the synthesis of supervisors subject to ω -language specifications*, in Proceedings of 22nd Annual Conference on Information Sciences and

- Systems, Princeton, NJ, 1988, pp. 440–444.
- [23] S. YOUNG, D. SPANJOL, AND V. K. GARG, *Control of discrete event systems modeled with infinite strings*, in Proceedings of 1992 American Control Conference, Chicago, IL, 1992, pp. 2809–2813.

ASYMPTOTIC STABILITY OF INFINITE-DIMENSIONAL DISCRETE-TIME BALANCED REALIZATIONS*

RAIMUND OBER[†] AND YUANYIN WU[‡]

Abstract. The question of power and asymptotic stability of infinite-dimensional discrete-time state space systems is investigated. It is shown that every balanced realization is asymptotically stable. Conditions are given for balanced, input normal, or output normal realizations to be asymptotically and/or power stable.

Key words. linear infinite-dimensional systems, balanced realizations, stability, Hankel operator

AMS subject classifications. 93B15, 93B20, 93B28, 93D20

1. Introduction. Balanced realizations for finite-dimensional systems have received a great deal of attention. They were introduced as a means of performing model reduction in an easy fashion [7] and have subsequently been used in H^∞ control theory, for example, to evaluate the Hankel norm of a linear system [4], [5]. Recently, they have been used to study parametrization problems of certain sets of linear systems [9].

The elegant results obtained for finite-dimensional balanced systems aroused interest in the problem of the extension of the notion of a balanced realization to infinite-dimensional systems. Glover, Curtain, and Partington [5] derived continuous-time balanced realizations for a class of systems with nuclear Hankel operators. Young [13] developed a very general realization theory for infinite discrete-time systems. Similar results were obtained in the continuous-time case by Ober and Montgomery-Smith [10].

One of the fundamental problems in systems theory is the question of stability of the system. In this paper, we will address this problem in the case of infinite-dimensional balanced realizations and the closely related input and output normal realizations. We show that every balanced realization is asymptotically stable. In general, input normal and output normal realizations do not have the same stability properties as balanced realizations, but we can also give necessary and sufficient conditions for them to be asymptotically and/or power stable. The result is that an input normal or output normal realization is power stable if and only if its transfer function is rational and proper with poles inside the open unit disk, whereas the power stability of a parbalanced realization is more complicated to characterize in terms of the properties of the transfer function.

The approach we take in the proofs of the results is to relate balanced realizations and, in particular, the input and output normal realizations to restricted shift realizations. We start in §2 with the restricted and *-restricted shift realizations and study their connections with Hankel operators, shift operators, and the Douglas–Shapiro–Shields factorizations of analytical functions. In §3, using these connections and the spectral theory of shift operators, we are able to give the above-mentioned necessary and sufficient conditions for the asymptotic and power stability of the output normal

* Received by the editors November 1, 1991; accepted for publication (in revised form) May 20, 1992.

[†] Center for Engineering Mathematics, University of Texas at Dallas, Richardson, Texas 75083-0688.

[‡] Center for Engineering Mathematics, University of Texas at Dallas, Richardson, Texas 75083-0688. This research was supported by grant 00974103 from Texas Advanced Research Program.

and input normal realizations. Young [13] established the existence of parbalanced realization for any function $G \in TLD^{U,Y}$. In §4 we prove that parbalanced realizations are always asymptotically stable. We also give examples that show the difficulty to analyze the power stability of a parbalanced realization in connection with its transfer function. A concluding remark is given on how to restrict ourselves to a slightly smaller class of discrete-time transfer functions and linear systems so that we can transpose all our results to the continuous-time case using a bilinear mapping (see [10]).

The following symbols are used:

\mathbb{D}	the open unit disk,
$\partial\mathbb{D}$	the unit circle,
\mathbb{D}_e	the exterior of $(\partial\mathbb{D}) \cup \mathbb{D}$,
$D_X^{U,Y}$	defined in §2,
$G^\perp(z)$	$(1/z)[G(1/z) - G(\infty)]$, $z \in \mathbb{D}$ for $G \in TLD^{U,Y}$,
H_K	the Hankel operator with symbol K ,
$H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$	$\{F \mid F : \mathbb{D} \rightarrow \mathcal{L}(U, Y) \text{ analytic and bounded on } \mathbb{D}\}$,
$H_Y^2(\mathbb{D})$	$\{f \mid f : \mathbb{D} \rightarrow Y \text{ analytic on } \mathbb{D} \text{ and}$ $\sup_{0 < r < 1} \int_0^{2\pi} \ f(re^{it})\ ^2 dt < \infty\}$,
J	$L_U^2(\partial\mathbb{D}) \rightarrow L_Y^2(\partial\mathbb{D})$, $(Jf)(z) = f(z^{-1})$,
$\tilde{K}(z)$	$(K(\bar{z}))^*$,
$\mathcal{L}(U, Y)$	$\{A \mid A : U \rightarrow Y \text{ a bounded operator}\}$,
$L_Y^2(\partial\mathbb{D})$	$\{f \mid f : \partial\mathbb{D} \rightarrow Y \text{ square integrable on } \partial\mathbb{D}\}$,
$L_{\mathcal{L}(U,Y)}^\infty(\partial\mathbb{D})$	$\{F \mid F : \partial\mathbb{D} \rightarrow \mathcal{L}(U, Y) \text{ measurable and essentially bounded}$ on $\partial\mathbb{D}\}$,
P_+	the orthogonal projection of $L_Y^2(\partial\mathbb{D})$ onto $H_Y^2(\mathbb{D})$,
P_X	the orthogonal projection of $H_Y^2(\mathbb{D})$ onto $X \subseteq H_Y^2(\mathbb{D})$,
S	the forward shift: $(Sf)(z) = zf(z)$ for $f \in H_Y^2(\mathbb{D})$,
S^*	the backward shift: $(S^*f)(z) = z^{-1}[f(z) - f(0)]$ for $f \in H_Y^2(\mathbb{D})$,
$S(Q)$	$P_X S _X$, the compression of S to X ,
	where $X = H_Y^2(\mathbb{D}) \ominus (QH_Y^2(\mathbb{D}))$,
$S(Q)^*$	$S^* _{H_Y^2(\mathbb{D}) \ominus (QH_Y^2(\mathbb{D}))}$, the restriction of S^* to $H_Y^2(\mathbb{D}) \ominus (QH_Y^2(\mathbb{D}))$,
$\sigma(A)$	the spectrum of an operator A ,
$\sigma_p(A)$	the point spectrum of an operator A ,
$\sigma(Q)$	the spectrum of an inner function $Q \in H_Y^\infty(\mathbb{D})$ (see §3),
$\sigma_s(G)$	the set of points in \mathbb{C} where G has no analytic continuation (see Theorem 3.14),
$TLD^{U,Y}$	defined in §3,
$X \vee Y$	closed linear span of subsets X and Y of a Hilbert space,
$(F, G)_L = I_Y$	F and G are weakly left coprime (see §2),
$(F, G)_R = I_U$	F and G are weakly right coprime (see §2).

2. Hankel operators and shift realizations for discrete-time systems.

Our results will be based on the analysis of restricted shift realizations whereby the shift realizations can be analyzed in terms of Hankel operators related to the transfer functions. Here we give a brief summary of some results on Hankel operators and the restricted shift realizations of discrete-time transfer functions. We start with some basic definitions.

Let U , X , and Y be separable Hilbert spaces. The linear systems considered in this paper are of the following form:

$$\left. \begin{aligned} x_{k+1} &= Ax_k + Bu_k, \\ y_k &= Cx_k + Du_k, \end{aligned} \right\} k = 0, 1, \dots,$$

where $u_k \in U$, $x_k \in X$, and $y_k \in Y$. The system operators are assumed to be such that A is a contraction on X , $B \in \mathcal{L}(U, X)$, $C \in \mathcal{L}(X, Y)$, and $D \in \mathcal{L}(U, Y)$. This system will be denoted by (A, B, C, D) and the set of all such systems $D_X^{U,Y}$. Unless otherwise stated, the spaces U , X , and Y are assumed to be infinite-dimensional.

For $(A, B, C, D) \in D_X^{U,Y}$, the function $G(z) = C(zI - A)^{-1}B + D$ is called the *transfer function* of (A, B, C, D) and (A, B, C, D) is called a *realization* of G . The *observability operator* $\mathcal{O} : D(\mathcal{O}) \rightarrow H_Y^2(\mathbb{D})$ of the system (A, B, C, D) is defined as

$$(\mathcal{O}x)(z) = \sum_{k \geq 0} (CA^k x)z^k$$

for $x \in D(\mathcal{O}) := \{x \in X \mid \sum_{k \geq 0} (CA^k x)z^k \in H_Y^2\}$. If $D(\mathcal{O}) = X$, \mathcal{O} is bounded and $\text{Ker}(\mathcal{O}) = \{0\}$, then the system (A, B, C, D) is said to be *observable*. The *dual system* of (A, B, C, D) is defined to be (A^*, C^*, B^*, D^*) , which is, in fact, a realization of the transfer function $\tilde{G}(z) := (G(\bar{z}))^*$. The system (A, B, C, D) is said to be *reachable* if its dual system is observable, and the *reachability operator* \mathcal{R} of (A, B, C, D) is defined to be the adjoint of the observability operator of the dual system. In fact, (A, B, C, D) is reachable if and only if the range of $\mathcal{R} : H_U^2(\mathbb{D}) \rightarrow X$ is dense in X , and in this case

$$\mathcal{R} \left(\sum_{k \geq 0} u_k z^k \right) = \sum_{k \geq 0} A^k B u_k, \quad \left(\sum_{k \geq 0} u_k z^k \in H_U^2 \right).$$

Note that we define the observability operator \mathcal{O} to have range in $H_Y^2(\mathbb{D})$ instead of l_Y^2 . Accordingly the domain of the reachability operator \mathcal{R} is in $H_U^2(\mathbb{D})$ instead of l_U^2 . The definitions adapted here are found to be more convenient in our context.

We write $LD_X^{U,Y}$ for the class of reachable and observable systems with state space X . The set of $\mathcal{L}(U, Y)$ -valued transfer functions that have reachable and observable realizations is denoted by $TLD^{U,Y}$. Note that $(A, B, C, D) \in LD_X^{U,Y}$ if and only if $(A^*, C^*, B^*, D^*) \in LD_X^{Y,U}$. Correspondingly, $G \in TLD^{U,Y}$ if and only if $\tilde{G} \in TLD^{Y,U}$, where $\tilde{G}(z) = (G(\bar{z}))^*$, $(z \in \mathbb{D}_e)$.

For an observable and reachable system (A, B, C, D) with observability operator \mathcal{O} and reachability operator \mathcal{R} , the *observability gramian* is defined to be $\mathcal{M} := \mathcal{O}^* \mathcal{O} : X \rightarrow X$, and the *reachability gramian* is $\mathcal{W} := \mathcal{R} \mathcal{R}^* : X \rightarrow X$. If $\mathcal{M} = \mathcal{W}$, then the system is said to be *parbalanced*.

Let G be in $TLD^{U,Y}$; i.e., G has a reachable and observable realization $(A, B, C, D) \in LD_X^{U,Y}$ for some state space X . Let \mathcal{R} be the reachability operator and \mathcal{O} the observability operator of the realization. Hence the operator $\mathcal{OR} : H_U^2(\mathbb{D}) \rightarrow H_Y^2(\mathbb{D})$ is

bounded. By the fact that

$$G^\perp(z) = z^{-1}[G(z^{-1}) - G(\infty)] = C(I - zA)^{-1}B = \sum_{n \geq 0} CA^n Bz^n, \quad z \in \mathbb{D},$$

it can be verified that, for any polynomial $f(z) = \sum_{k=0}^n u_k z^k$, $u_k \in U$,

$$\mathcal{O}Rf = P_+ G^\perp Jf,$$

where $(Jg)(z) = g(z^{-1})$ for any $g \in H_U^2(\mathbb{D})$. In this way $P_+ G^\perp J : H_U^2(\mathbb{D}) \rightarrow H_Y^2(\mathbb{D})$ defines a bounded operator. It is called the *Hankel operator with symbol G^\perp* and is denoted by H_{G^\perp} .

Conversely, let G be a $\mathcal{L}(U, Y)$ -valued function, analytic on \mathbb{D}_e and at infinity such that the Hankel operator $H_{G^\perp} = P_+ G^\perp J : H_U^2(\mathbb{D}) \rightarrow H_Y^2(\mathbb{D})$ is defined for every polynomial $f(z) = \sum_{k=0}^n u_k z^k$, ($u_k \in U$) and can be extended to a bounded operator. Then G has reachable and observable realizations. In fact, G has the restricted shift realization, which was first introduced by Fuhrmann [2] and Helton [6] (see also [13]).

THEOREM 2.1. *Let G be a $\mathcal{L}(U, Y)$ -valued function analytic on \mathbb{D}_e and at infinity such that $H_{G^\perp} : H_U^2(\mathbb{D}) \rightarrow H_Y^2(\mathbb{D})$ defines a bounded operator. Then G has a state space realization (A, B, C, D) with state space X , i.e., for $z \in \mathbb{D}_e$,*

$$G(z) = C(zI - A)^{-1}B + D,$$

which is given in the following way:

The state space X is given by

$$X = \overline{\text{range}} H_{G^\perp} \subseteq H_Y^2(\mathbb{D}).$$

The state propagation operator $A : X \rightarrow X$, the input operator $B : U \rightarrow X$, the output operator $C : X \rightarrow Y$ and the feedthrough operator $D : U \rightarrow Y$ are given by the following, for $f \in X$ and $u \in U$:

$$\begin{aligned} (Af)(z) &:= (S^*f)(z) = \frac{f(z) - f(0)}{z}, \\ (Bu)(z) &:= G^\perp(z)u, \\ Cf &:= f(0), \\ Du &:= G(\infty)u, \end{aligned}$$

where S is the (forward) shift operator: $(Sf)(z) = zf(z)$, $f \in H_Y^2(\mathbb{D})$. The realization (A, B, C, D) is called the restricted shift realization of the transfer function G .

The following proposition shows that the restricted shift realization is reachable and observable.

PROPOSITION 2.2 (see [3] or [13]). *Assume the notation of Theorem 2.1. Then the system (A, B, C, D) is in $LD_X^{U,Y}$; i.e., it is observable and reachable. The observability operator \mathcal{O} and reachability operator \mathcal{R} of (A, B, C, D) are, respectively, given by*

$$\mathcal{O} = I_X : X \rightarrow H_Y^2(\mathbb{D}) \quad \text{and} \quad \mathcal{R} = H_{G^\perp} : H_U^2(\mathbb{D}) \rightarrow X.$$

Therefore the class $TL D^{U,Y}$ of transfer functions can be characterized as the set of $\mathcal{L}(U, Y)$ -valued functions analytic on \mathbb{D}_e and at infinity such that the Hankel operator H_{G^\perp} is bounded. For such transfer functions, the restricted shift realization exists. We emphasize these points by the following corollary.

COROLLARY 2.3. *The following statements are equivalent:*

1. G is in $TLD^{U,Y}$, i.e. G has a reachable and observable realization in some state space X ;
2. G has the restricted shift realization that is reachable and observable;
3. G is analytic on \mathbb{D}_e and at infinity such that the Hankel operator H_{G^\perp} is bounded.

Note that, if $G^\perp \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$, then $G \in TLD^{U,Y}$, since, in this case, H_{G^\perp} is bounded.

As a next step, we construct another realization, which is the dual realization of the restricted shift realization. Let G be in $TLD^{U,Y}$. Then $\tilde{G}(z) = (G(\bar{z}))^*$, ($z \in \mathbb{D}_e$) is in $TLD^{Y,U}$. Moreover, if $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is the restricted shift realization of \tilde{G} , then the dual system $(\tilde{A}^*, \tilde{C}^*, \tilde{B}^*, \tilde{D}^*)$ is a realization of G , called the **-restricted shift realization* of G .

A concrete representation of the *-restricted shift realization can be obtained.

THEOREM 2.4. *Let G be in $TLD^{U,Y}$. The state space representation (A_*, B_*, C_*, D_*) of the *-restricted shift realization is given by the following:*

The state space X_ is $X_* = \overline{\text{range}} H_{\tilde{G}^\perp}$, where*

$$\tilde{G}^\perp(z) = (G^\perp(\bar{z}))^*.$$

The operators A_, B_*, C_* , and D_* are defined as*

$$\begin{aligned} A_* &= P_{X_*} S|_{X_*}, \\ B_* u &= P_{X_*} u, \quad (u \in U) \\ C_* f &= (H_{G^\perp} f)(0), \quad (f \in X_*), \\ D_* &= G(\infty), \end{aligned}$$

where P_{X_} is the orthogonal projection of $H_U^2(\mathbb{D})$ onto X_* , and the space U is considered as the subspace $\{u + 0z + 0z^2 + 0z^3 + \dots \mid u \in U\}$ of $H_U^2(\mathbb{D})$.*

The system (A^, C^*, B^*, D^*) is observable and reachable. The reachability and observability operators \mathcal{R}_* and \mathcal{O}_* are, respectively, given by*

$$\mathcal{R}_* = P_{X_*} : H_U^2(\mathbb{D}) \rightarrow X_* \quad \text{and} \quad \mathcal{O}_* = H_{\tilde{G}^\perp}^*|_{X_*} = H_{G^\perp}|_{X_*}.$$

Proof. Replacing G by \tilde{G} in Theorem 2.1, we obtain the restricted shift realization of \tilde{G} , and the dual of this realization is the *-restricted shift realization stated in the theorem. Here we just verify the formula for the output operator C_* , which is the adjoint of the input operator \tilde{B} of the restricted shift realization of \tilde{G} . Hence $C_* = \tilde{B}^*$. So by Theorem 2.1 we have

$$C_*^* y = \tilde{B} y = \tilde{G}^\perp y \in X_*, \quad y \in Y.$$

From this, we obtain that, for $f \in X_* \subseteq H_U^2(\mathbb{D})$ and $y \in Y$,

$$\begin{aligned} \langle C_* f, y \rangle_Y &= \langle f, C_*^* y \rangle_{H_U^2(\mathbb{D})} \\ &= \frac{1}{2\pi} \int_0^{2\pi} \langle f(e^{it}), \tilde{G}^\perp(e^{it}) y \rangle_U dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} \langle (\tilde{G}^\perp(e^{it}))^* f(e^{it}), y \rangle_U dt \\ &= \langle \frac{1}{2\pi} \int_0^{2\pi} (\tilde{G}^\perp(e^{it}))^* f(e^{it}) dt, y \rangle_U. \end{aligned}$$

Hence, using a change of variable $z = e^{-it}$, we have

$$\begin{aligned} C_*f &= \frac{1}{2\pi} \int_0^{2\pi} \left(\tilde{G}^\perp(e^{it}) \right)^* f(e^{it}) dt \\ &= \frac{1}{2\pi i} \int_{\partial\mathbb{D}} \bar{z} (\tilde{G}^\perp(\bar{z}))^* f(\bar{z}) dz \\ &= \frac{1}{2\pi i} \int_{\partial\mathbb{D}} \bar{z} G^\perp(z) f(\bar{z}) dz. \end{aligned}$$

Note that the last integral is the zeroth Fourier coefficient of $G^\perp(z)f(\bar{z})$, which is the same as the zeroth Fourier coefficient of $P_+G^\perp(z)f(\bar{z}) = H_{G^\perp}f$. This is $(H_{G^\perp}f)(0)$, since $H_{G^\perp}f \in H_Y^2(\mathbb{D})$. \square

From these results, we see that the state space for the restricted shift realization is given as the closed range of the Hankel operator whose symbol is the transfer function mapped to the unit disk. The state propagation operator is just the backward shift restricted to the state space. For the $*$ -restricted shift realization, the state space is also the closed range of a Hankel operator, while the state propagation operator is the forward shift compressed to this state space. It is well known and readily verified that the closure of the range of a Hankel operator H_G is the orthogonal complement of a right invariant subspace of $H_Y^2(\mathbb{D})$ [3], [8]. A vector-valued version of Beurling's theorem (see, e.g., [3, Thm. 12.22, p. 186]) asserts that a right invariant space in $H_Y^2(\mathbb{D})$ can only be either the trivial space $\{0\}$ or $QH_Y^2(\mathbb{D})$, where $Q \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$ is such that $\|Q\|_\infty \leq 1$ and $Q(e^{it})$ is for almost every $t \in [0, 2\pi)$ a partial isometry with a fixed nonzero initial space. Such a function Q is called a *rigid* function. A rigid function Q is called *inner* if $Q(e^{it})$ is a unitary operator for almost all $t \in [0, 2\pi)$.

This discussion leads to the cyclicity of functions defined as follows (see [3]).

DEFINITION 2.1. Let $G \in TLD^{U,Y}$. Then G^\perp is called

1. *cyclic* if $(\text{range } H_{G^\perp})^\perp = \{0\}$,
2. *noncyclic* if $(\text{range } H_{G^\perp})^\perp = QH_Y^2(\mathbb{D})$ for some rigid function $Q \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$,
3. *strictly noncyclic* if $(\text{range } H_{G^\perp})^\perp = QH_Y^2(\mathbb{D})$ for some inner function $Q \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$.

It should be noted that the inner function Q in statement 3 of the above definition is unique up to right multiplication by a constant unitary operator on Y . Also, if G^\perp is scalar, then G^\perp is noncyclic if and only if it is strictly noncyclic. It is important to have characterizations for matrix-valued functions to be strictly noncyclic. To this end, we introduce some definitions. Let K be in $H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$. The function \hat{K} defined on \mathbb{D}_e with values in $\mathcal{L}(U, Y)$ is called a *meromorphic pseudocontinuation of bounded type* of K if \hat{K} is of bounded type, i.e.,

$$\hat{K}(z) = \frac{F(z)}{h(z)}, \quad z \in \mathbb{D}_e,$$

where F is a $\mathcal{L}(U, Y)$ -valued function and h is a scalar-valued function, both bounded and analytic in \mathbb{D}_e ; K and \hat{K} have the same strong radial limits on $\partial\mathbb{D}$.

Let $F_1 \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and $F_2 \in H_{\mathcal{L}(Z,Y)}^\infty(\mathbb{D})$. We say that F_1 and F_2 are *left weakly coprime* and write

$$(F_1, F_2)_L = I_Y$$

if $F_1 H_U^2(\mathbb{D}) \vee F_2 H_Z^2(\mathbb{D}) = H_Y^2(\mathbb{D})$, where \vee stands for the closed linear span.

Analogously, we say that $K_1 \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and $K_2 \in H_{\mathcal{L}(U,Z)}^\infty(\mathbb{D})$ are *weakly right coprime* and write $(K_1, K_2)_R = I_U$ if \tilde{K}_1 and \tilde{K}_2 are weakly left coprime.

Using these notations, we have the following theorem ([3, Thm. 3.5, p. 254]).

THEOREM 2.5. *For $K \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ with U and Y finite-dimensional, the following statements are equivalent:*

1. K is strictly noncyclic,
2. On $\partial\mathbb{D}$ the function K can be factored as

$$K = Q_1(zF_1)^* = (zF_2)^*Q_2.$$

Q_1 and Q_2 are inner functions in $H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$ and $H_{\mathcal{L}(U)}^\infty(\mathbb{D})$, respectively. The functions F_1 and F_2 are in $H_{\mathcal{L}(Y,U)}^\infty(\mathbb{D})$ and in $H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$, respectively, and the coprimeness conditions $(Q_1, F_1)_R = I_Y$, $(Q_2, F_2)_L = I_U$ hold. Here Q_1 (respectively, Q_2) is unique up to right (respectively, left) multiplication by a constant unitary operator,

3. K has a meromorphic pseudocontinuation of bounded type on \mathbb{D}_e .

If statement 2 holds, then $Q_1H_Y^2(\mathbb{D}) = (\text{range } H_K)^\perp$ and $\tilde{Q}_2H_U^2(\mathbb{D}) = (\text{range } H_{\tilde{K}})^\perp$.

We will call the factorization of K in the theorem the Douglas–Shapiro–Shields factorization. In fact, this is the generalization due to Fuhrmann [3] of the result on scalar functions of Douglas, Shapiro, and Shields [1].

By Theorem 2.5, we immediately have the following corollary.

COROLLARY 2.6. *In the notation of the theorem with U and Y finite-dimensional, K is strictly noncyclic if and only if \tilde{K} is strictly noncyclic.* \square

From Theorems 2.1, 2.4, 2.5, and Definition 2.1, we see that the state space of a restricted shift realization of a transfer function G is the orthogonal complement of an invariant subspace, which is characterized by a rigid function Q . The state propagation operator A is the backward shift S^* restricted to the state space $(QH_Y^2(\mathbb{D}))^\perp$, i.e., $A = S^*_{|(QH_Y^2(\mathbb{D}))^\perp}$, which we will denote by $S(Q)^*$. One of the important points in our context is that the function Q can be determined from the transfer function G , if G^\perp is strictly noncyclic.

For the $*$ -restricted shift realization, the state space can be determined in a similar way to the derivation of the restricted shift realization. In this case, the state propagation operator is the forward shift operator S compressed to the orthogonal complement of an invariant subspace that is determined by a rigid function Q_* , i.e., $P_{(Q_*H_U^2(\mathbb{D}))^\perp}S|_{(Q_*H_U^2(\mathbb{D}))^\perp}$, which we denote by $S(Q_*)$.

We summarize these results in the following proposition.

PROPOSITION 2.7. *Let G be in $TLD^{U,Y}$ with U and Y finite-dimensional and let $(A, B, C, D) \in LD_X^{U,Y}$ be its restricted shift realization and $(A_*, B_*, C_*, D_*) \in LD_{X_*}^{U,Y}$ its $*$ -restricted shift realization. Then*

1. If G^\perp is cyclic we have that (a) $A = S^*$ and $X = H_Y^2(\mathbb{D})$, and (b) $A_* = S$ and $X_* = H_U^2(\mathbb{D})$;
2. If G^\perp is noncyclic, we have that (a) $A = S(Q)^*$, where $Q \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$ is a rigid function such that

$$X = \overline{\text{range}} H_{G^\perp} = (QH_Y^2(\mathbb{D}))^\perp.$$

If G^\perp is in $H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and is strictly noncyclic with factorization $G^\perp = Q_1(zF_1)^*$, where $Q_1 \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$ is inner and $F_1 \in H_{\mathcal{L}(Y,U)}^\infty(\mathbb{D})$ such that $(Q_1, F_1)_R = I_Y$, then $Q = Q_1V_1$ for some unitary operator V_1 on Y , and (b) $A_* = P_{X_*}S|_{X_*} = S(Q_*)$, where $Q_* \in H_{\mathcal{L}(U)}^\infty(\mathbb{D})$ is a rigid function such that

$$X_* = \overline{\text{range}} H_{\tilde{G}^\perp} = (Q_*H_U^2(\mathbb{D}))^\perp.$$

If G^\perp is in $H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and is strictly noncyclic and \tilde{G}^\perp has a factorization $\tilde{G}^\perp = Q_2(zF_2)^*$, where $Q_2 \in H_{\mathcal{L}(U)}^\infty(\mathbb{D})$ is inner and $F_2 \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ such that $(Q_2, F_2)_L = I_U$, then $Q_* = Q_2 V_2$ for some unitary operator V_2 on U .

Proof. The proposition follows from Theorems 2.1, 2.4, 2.5, and Definition 2.1.

□

3. Stability and spectral minimality of input normal and output normal realizations. In this section, we discuss the stability and questions of spectral minimality of input normal and output normal realizations using the results on restricted and *-restricted shift realizations studied in §2.

The following definition recalls the notion of an input normal and output normal system as defined by Moore [7] for finite-dimensional state space realizations. The definitions in the infinite-dimensional case are natural extensions of the finite-dimensional notions (see, e.g., [13]).

DEFINITION 3.1. Let (A, B, C, D) be in $LD_X^{U,Y}$. Then the system is

1. *output normal* if $\mathcal{M} = I$,
2. *input normal* if $\mathcal{W} = I$,
3. *parbalanced* if $\mathcal{M} = \mathcal{W}$,

4. *balanced* if $\mathcal{M} = \mathcal{W}$ and there is an orthonormal basis of the state space with respect to which \mathcal{M} (and hence \mathcal{W}) has a diagonal matrix representation.

From our results on the restricted and the *-restricted shift realization we immediately have examples for input and output normal realizations.

PROPOSITION 3.1. Let $G \in TLD^{U,Y}$. Then the restricted shift realization is output normal whereas the *-restricted shift realization is input normal.

Proof. The proof follows from Proposition 2.2 and Theorem 2.4 □

Next, we quote a result that establishes a reachable output-normal realization of a transfer function is unitarily equivalent to its restricted shift realization

Two systems $(A_1, B_1, C_1, D_1) \in D_{X_1}^{U,Y}$ and $(A_2, B_2, C_2, D_2) \in D_{X_2}^{U,Y}$ are called *equivalent* (*unitarily equivalent*) if there exists a bounded and boundedly invertible operator (a unitary operator) V mapping the state space X_1 onto the state space X_2 , such that

$$(A_1, B_1, C_1, D_1) = (V^{-1}A_2V, V^{-1}B_2, C_2V, D_2).$$

In this case, V is called an equivalence (unitary) transformation.

THEOREM 3.2 (see [13]). Let $(A_1, B_1, C_1, D_1) \in LD_{X_1}^{U,Y}$ and $(A_2, B_2, C_2, D_2) \in LD_{X_2}^{U,Y}$ be two output normal realizations of a transfer function in $TLD^{U,Y}$. Then (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) are unitarily equivalent.

By a duality argument, we have as a corollary that the same result holds for input normal realizations; i.e., an input normal realization is unitarily equivalent to the *-restricted shift realization.

COROLLARY 3.3. Let $(A_1, B_1, C_1, D_1) \in LD_{X_1}^{U,Y}$ and $(A_2, B_2, C_2, D_2) \in LD_{X_2}^{U,Y}$ be two input normal realizations of a transfer function in $TLD^{U,Y}$. Then (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) are unitarily equivalent.

We now turn to the study of stability. We introduce a classification of contractions according to their stability properties [12], which will simplify our notation.

DEFINITION 3.2. Let T be a contraction on the Hilbert space H . Then

1. $T \in C_0$ if $\lim_{n \rightarrow \infty} T^n h = 0$, for all $h \in H$,
2. $T \in C_0$ if $\lim_{n \rightarrow \infty} (T^*)^n h = 0$, for all $h \in H$,
3. $T \in C_1$ if $\lim_{n \rightarrow \infty} T^n h \neq 0$, for all $h \in H$, $h \neq 0$,

4. $T \in C_{\cdot 1}$ if $\lim_{n \rightarrow \infty} (T^*)^n h \neq 0$, for all $h \in H$, $h \neq 0$.

We further set $C_{ij} = C_{\cdot i} \cap C_{\cdot j}$, $i, j = 0, 1$.

Now we define the two notions of stability we will consider in the remainder of the paper.

DEFINITION 3.3. A discrete time system $(A, B, C, D) \in D_X^{U, Y}$ or the state propagation operator A is called

1. *asymptotically stable* if for every $x \in X$,

$$A^k x \rightarrow 0 \quad \text{as } k \rightarrow \infty,$$

i.e., if A is of class $C_{0\cdot}$,

2. *power stable* if $r < 1$, where

$$r := \inf\{\bar{r} \mid \text{there is } M_{\bar{r}} > 0 \text{ such that } \|A^k\| \leq M_{\bar{r}} \bar{r}^k, k \geq 0\}.$$

The number r is called the *degree of power stability*.

It is easy to see that stability and observability, as well as reachability properties of discrete time systems, are preserved under equivalence transformations, whereas input and output normality are preserved under unitary equivalence. Moreover, two equivalent power stable systems have the same degree of power stability.

Therefore, by Theorem 3.2 and its corollary, we can establish all stability and other important results concerning input normal and output normal realizations by restricting ourselves to $*$ -restricted and restricted shift realizations. Henceforth, when we prove statements about input normal or output normal reachable and observable realizations, we must only prove them in the case of restricted or $*$ -restricted realization.

From Proposition 2.7, we can see that the study of stability and spectral properties of the restricted and $*$ -restricted realizations reduces to the study of the operators $S(Q)^*$ and $S(Q_*)$, where Q and Q_* are rigid functions. We will need the following lemma (see [8, Cor., p. 43]).

LEMMA 3.4. Let $Q \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$ be a rigid function. Denote by P_X the projection on $X := (QH_Y^2(\mathbb{D}))^\perp$. Then, for $f \in H_Y^2(\mathbb{D})$, $\lim_{n \rightarrow \infty} \|P_X S^n f\|^2 = \|f\|^2 - \|Q^* f\|^2$.

The following theorem shows that an output normal realization of a transfer function in $TLD^{U, Y}$ is always asymptotically stable.

THEOREM 3.5. Let $G \in TLD^{U, Y}$ and let (A, B, C, D) be an output normal reachable realization of G . Then

1. $A \in C_{0\cdot}$; i.e., A is asymptotically stable,
2. $A \in C_{00}$ if G^\perp is strictly noncyclic,
3. $A \in C_{01}$ if G^\perp is cyclic.

Proof. By Proposition 3.1, we can assume without loss of generality that (A, B, C, D) is the restricted shift realization.

1. The state propagation operator A of the restricted shift realization is the restriction of the backward shift to a subspace of $H_Y^2(\mathbb{D})$. The backward shift S^* is such that for every $x_0 \in H_Y^2(\mathbb{D})$,

$$(S^*)^k x_0 \rightarrow 0, \quad \text{as } k \rightarrow \infty.$$

This immediately implies statement 1.

2. This follows from Proposition 2.7 and Lemma 3.4.

3. If G^\perp is cyclic, then A is the backward shift S^* on the space $H_Y^2(\mathbb{D})$, and therefore $A \in C_{01}$. \square

In the case of input normal realizations, the situation is, however, such that we cannot, in general, expect that the realization is asymptotically stable, since the state propagation operator of the $*$ -restricted shift realization is the forward shift operator, compressed to a subspace of $H_U^2(\mathbb{D})$. The forward shift on $H_U^2(\mathbb{D})$ is not asymptotically stable. The following corollary states that, at least for an important class of transfer functions, input normal realizations are asymptotically stable.

COROLLARY 3.6. *Let $G \in \text{TLD}^{U,Y}$ and let (A, B, C, D) be an input normal observable realization of G . Then*

1. $A \in C_0$,
2. $A \in C_{00}$ if \tilde{G}^\perp is strictly noncyclic,
3. $A \in C_{10}$ if \tilde{G}^\perp is cyclic.

Proof. Let (A, B, C, D) be the $*$ -restricted realization of G . Recall that by definition (A, B, C, D) is the dual system of the restricted shift realization of \tilde{G} . Hence the result follows by duality from Theorem 3.5 \square

We now proceed to power stability. The following result gives a characterization of power stability (see, e.g., Przyluski [11]).

PROPOSITION 3.7. *Let T be a contraction. Then the spectral radius $r(T)$ of T , i.e.,*

$$r(T) = \sup\{|\lambda| \mid \lambda \in \sigma(T)\},$$

is given by

$$r(T) = \inf\{0 \leq \bar{r} \leq 1 \mid \text{there exists } M_{\bar{r}} \geq 0 \text{ such that } \|T^k\| \leq M_{\bar{r}} \bar{r}^k, k \geq 0\}.$$

Hence, if T is power-stable, then the degree of power stability equals the spectral radius.

Proof. The proof follows from an application of the well-known formula

$$\sup\{|\lambda| \mid \lambda \in \sigma(T)\} = \lim_{n \rightarrow \infty} \|T^n\|^{1/n}. \quad \square$$

To establish whether the output normal and input normal realizations are power-stable, it is therefore important to determine the spectral radius of its state propagation operator. To this end, we must introduce C_0 contractions, which play an important role in the theory of contractive operators. C_0 contractions are defined via the H^∞ calculus for contractions and are a special class of completely nonunitary contractions (see [12]). Specifically, a contraction T on a Hilbert space H is *completely nonunitary* if there is no subspace $V \subseteq H$ such that $TV = V$ and $T|_V$ is unitary. For such T , the operator $u(T) := \lim_{r \rightarrow 1} u(rT)$ is a well-defined bounded operator for any $u \in H^\infty$ and satisfies $\|u(T)\| \leq \|u\|_{H^\infty}$. In particular, $u(T)$ is a contraction if u is an inner function.

A completely nonunitary contraction is a C_0 contraction if there exists an inner function m such that $m(T) = 0$. The least common divisor of all such inner functions is called the *minimal function* m_T of T . For the minimal function m_T , we also have that $m_T(T) = 0$. Therefore the minimal function of a C_0 contraction can be seen to be a generalization of the minimal polynomials for matrices.

As in the case of matrices, the spectrum of C_0 operators is given by the “zeros” of the minimal function in the following sense. We define the *spectrum* $\sigma(Q)$ of an inner function $Q \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$ to be

$$\sigma(Q) = \left\{ \lambda \in \overline{\mathbb{D}} \mid \lim_{\substack{\delta > 0 \\ \delta \rightarrow 0}} \inf_{\substack{\xi \in \mathbb{D} \\ |\xi - \lambda| < \delta}} \inf_{\substack{\|y\|=1 \\ y \in Y}} \|Q(\xi)y\| = 0 \right\}.$$

Then we have the following proposition (see [8, p. 75]),

PROPOSITION 3.8. *If T is a C_0 operator, then*

$$\sigma(T) = \sigma(m_T) \quad \text{and} \quad \sigma_p(T) = \sigma(m_T) \cap \mathbb{D}.$$

Now we use these results to analyze the spectrum of the operators $S(Q)$ and $S(Q)^*$. First the following proposition [8, p. 73] shows when $S(Q)$ is a C_0 contraction.

PROPOSITION 3.9. *If $\dim(U) < \infty$ and $Q \in H_{\mathcal{L}(U)}^\infty(\mathbb{D})$ is a rigid function, then the determinant $d = \det(Q)$ is such that $d(S(Q)) = 0$. Therefore, when Q is an inner function and U has finite dimension, the operator $S(Q)$ is a C_0 contraction.*

In fact, $S(Q)$ and $S(Q)^*$ are both C_0 contractions when Q is inner and U is finite dimensional, as shown by the following result (see [3, Thm, 13.2, p. 191] or [8, p. 75]).

PROPOSITION 3.10. *For a given inner function $Q \in H_{\mathcal{L}(U)}^\infty(\mathbb{D})$, the operators $S(Q)$ and $S(\tilde{Q})^*$ are unitarily equivalent.*

More precisely, $S(Q) = \tau_Q^{-1} S(\tilde{Q})^* \tau_Q$, where the unitary operator τ_Q is given by

$$\begin{aligned} \tau_Q : L_U^2(\partial\mathbb{D}) &\rightarrow L_U^2(\partial\mathbb{D}) \\ f &\mapsto e^{-it} \tilde{Q} J f. \end{aligned}$$

One of the important results in the theory of the backward shift operator $S(Q)^*$ restricted to an invariant subspace is that its spectrum can be completely characterized by the associated inner function Q . Note that, if σ is a set of complex numbers, then σ^* is used to denote the set of the complex conjugates of the elements in σ .

THEOREM 3.11 (see [8, p. 75]). *The following statements hold:*

1. (a) *Let S^* be the backward shift on $H_Y^2(\mathbb{D})$. Then*

$$\sigma(S^*) = \overline{\mathbb{D}}, \quad \sigma_p(S^*) = \mathbb{D},$$

- (b) *Let S be the forward shift on $H_Y^2(\mathbb{D})$. Then*

$$\sigma(S) = \overline{\mathbb{D}}, \quad \sigma_p(S) = \emptyset;$$

2. (a) *Let Q be an inner function in $H_Y^\infty(\mathbb{D})$ with Y finite dimensional. Then*

$$\sigma(S(Q)^*) = \sigma(Q)^* = \sigma(m_{S(Q)^*}),$$

$$\sigma_p(S(Q)^*) = \sigma(S(Q)^*) \cap \mathbb{D} = \{\bar{\lambda} \in \mathbb{D} \mid \text{Ker} Q(\lambda)^* \neq \{0\}\},$$

- (b)

$$\sigma(S(Q)) = \sigma(Q) = \sigma(m_{S(Q)}),$$

$$\sigma_p(S(Q)) = \sigma(S(Q)) \cap \mathbb{D} = \{\lambda \in \mathbb{D} \mid \text{Ker} Q(\lambda) \neq \{0\}\}.$$

The next result shows that we must only be concerned with inner functions if we are interested in the case when the spectral radius of the restricted backward shift is less than 1 (see [3, p. 194]).

THEOREM 3.12. *Let U be finite-dimensional and Q a rigid function that is not inner. Then $\sigma_p(S(Q)^*)$ is equal to the open unit disk \mathbb{D} .*

In terms of the restricted and *-restricted shift realizations, Theorems 3.11 and 3.12 can be translated into the following result.

PROPOSITION 3.13. *Let (A, B, C, D) and (A_*, B_*, C_*, D_*) be, respectively, the restricted and *-restricted shift realizations of a transfer function $G \in TLD^{U,Y}$ where U and Y have finite dimensions.*

1. *If G^\perp is cyclic, then $\sigma(A) = \overline{\mathbb{D}}$, $\sigma_p(A) = \mathbb{D}$ and $\sigma(A_*) = \overline{\mathbb{D}}$, and $\sigma_p(A) = \emptyset$.*
2. *If G^\perp is noncyclic but not strictly noncyclic, then $\sigma(A) = \overline{\mathbb{D}}$, $\sigma(A_*) = \overline{\mathbb{D}}$.*
3. *If G^\perp is in $H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and is strictly noncyclic with factorization $G^\perp = Q_1(zF_1)^*$ and $\tilde{G}^\perp = Q_2(zF_2)^*$, where $Q_1 \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$ and $Q_2 \in H_{\mathcal{L}(U)}^\infty(\mathbb{D})$ are inner, and where Q_1 and $F_1 \in H_{\mathcal{L}(Y,U)}^\infty(\mathbb{D})$ are right weakly coprime, and Q_2 and $F_2 \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ are also right weakly coprime, then*

$$\sigma(A) = \sigma(Q_1)^* = \sigma(m_A),$$

$$\sigma_p(A) = \sigma(Q_1)^* \cap \mathbb{D} = \{\bar{\lambda} \in \mathbb{D} \mid \text{Ker} Q_1(\lambda)^* \neq \{0\}\}$$

and

$$\sigma(A_*) = \sigma(Q_2) = \sigma(m_{A_*}),$$

$$\sigma_p(A_*) = \sigma(Q_2) \cap \mathbb{D} = \{\lambda \in \mathbb{D} \mid \text{Ker} Q_2(\lambda) \neq \{0\}\}.$$

Proof. The proposition follows from Theorems 3.11 and 3.12 and Proposition 2.7. \square

A very important property of finite-dimensional systems is that the eigenvalues of the state propagation matrix correspond exactly to the poles of the transfer function. For infinite-dimensional systems, it is desirable to have the analogous property. This was shown to be true for strictly noncyclic transfer functions by Fuhrmann ([3, Chap. III]).

DEFINITION 3.4. Let $G \in TLD^{U,Y}$ be such that G^\perp has a meromorphic pseudocontinuation of bounded type on \mathbb{D}_e . Then we extend the definition of G^\perp onto \mathbb{D}_e to be this unique meromorphic pseudocontinuation and hence define G on \mathbb{D} . The set $\sigma_s(G)$ is defined to be the set of points z such that the extended G cannot be analytically continued to z .

A realization $(A, B, C, D) \in LD_X^{U,Y}$ of G is said to be *spectrally minimal* if $\sigma(A) = \sigma_s(G)$.

We note that a more general definition of spectral minimality can be made for a larger class of transfer functions (see [3]). However, the definition suffices for our discussion here. It turns out that, if $G \in TLD^{U,Y}$ is strictly noncyclic, then both the restricted and *-restricted shift realizations are spectrally minimal.

THEOREM 3.14. *Let G be in $TLD^{U,Y}$, where U and Y have finite dimensions. If G^\perp is in $H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and is strictly noncyclic, then*

1. *Every output normal realization (A, B, C, D) is spectrally minimal, i.e.,*

$$\sigma(A) = \sigma_s(G),$$

2. *Every input normal realization (A_*, B_*, C_*, D_*) is spectrally minimal.*

Proof. 1. See, [3, Thm. 4.11, p. 267] for the case of the restricted shift realization.

2. Without loss of generality, we assume (A, B, C, D) is the $*$ -restricted shift realization of G .

Recall from Corollary 2.6 that G^\perp is strictly noncyclic if and only if \tilde{G}^\perp is strictly noncyclic. By statement 1 and the construction of the $*$ -restricted shift realization, we have that $\sigma(A_\star^*) = \sigma_s(\tilde{G})$. Since

$$\sigma_s(G) = (\sigma_s(\tilde{G}))^* = \sigma(A_\star^*)^* = \sigma(A_\star),$$

we have the spectral minimality of the $*$ -restricted shift realization. \square

We now show that an input normal or output normal system is power-stable if and only if it is finite-dimensional.

THEOREM 3.15. *Let G be in $TLD^{U,Y}$ such that $G^\perp \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and let U and Y be finite-dimensional. Then an output normal (respectively, input normal) realization of G is power-stable if and only if G is rational.*

Proof. Let G be rational. Since $G^\perp \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$, there is a number $0 < r < 1$ such that the poles of G are contained in the set $\{\lambda : |\lambda| \leq r\}$. Being rational, G^\perp has a meromorphic pseudocontinuation of bounded type and hence by Theorem 2.5 is strictly noncyclic. Theorem 3.14 then implies that the propagation operator of any output normal (respectively, input normal) realization of G has spectral radius less than 1. Now Proposition 3.7 shows that it is power-stable.

Conversely, assume that an output normal realization of G is power stable. Then the restricted shift realization is power stable. Let A be its propagation operator. We have by Proposition 3.7 that $r(A)$, the spectral radius of A , is less than 1. By Proposition 3.13, this implies that G^\perp is strictly noncyclic. Then, by Theorem 3.14, any output normal realization of G is spectrally minimal, and hence G can be analytically continued across $\partial\mathbb{D}$. Being strictly noncyclic, G^\perp has a meromorphic pseudocontinuation on \mathbb{D}_e , and thus G has a meromorphic pseudocontinuation on \mathbb{D} . Since a meromorphic pseudocontinuation is unique, the pseudocontinuation is an analytic continuation. Thus G is a meromorphic function on the extended complex plane. Hence it is rational.

If an input normal realization is assumed to be power-stable, a similar argument will also show that G is rational. \square

4. Balanced realizations. This section is devoted to the study of the stability properties of balanced realizations with infinite-dimensional state space.

Balanced realizations of finite-dimensional systems have played an important role in model reduction and Hankel norm approximation of linear systems [7], [4]. In finite dimensions, it is straightforward to construct a balanced realization from input normal or output normal realizations. In infinite dimensions, it is not trivial to guarantee that this can be done, since the state space transformation that is involved, in general, has an unbounded inverse. That this is nevertheless possible was shown by Young [13]. Note that, in the following theorem, the subscripts o and i signify output normal and input normal realizations, respectively.

THEOREM 4.1. *Let $G \in TLD^{U,Y}$. Let (A_o, B_o, C_o, D_o) be the restricted shift realization of G with state space $X_o = \overline{\text{range}} H_{G^\perp}$ and let (A_i, B_i, C_i, D_i) be the $*$ -restricted realization with state space $X_i = \overline{\text{range}} H_{\tilde{G}^\perp}$. Set*

$$\mathcal{W}_o = H_{G^\perp} H_{G^\perp}^*|_{X_o}, \quad \mathcal{M}_i = H_{\tilde{G}^\perp} H_{\tilde{G}^\perp}^*|_{X_i}.$$

1. There exist parbalanced realizations $(A_{b1}, B_{b1}, C_{b1}, D_{b1}) \in D_{X_o}^{U,Y}$ and $(A_{b2}, B_{b2}, C_{b2}, D_{b2}) \in D_{X_i}^{U,Y}$ of G that satisfy

$$\begin{aligned} \mathcal{W}_o^{1/4} A_{b1} &= A_o \mathcal{W}_o^{1/4}, & \mathcal{W}_o^{1/4} B_{b1} &= B_o, \\ C_{b1} &= C_o \mathcal{W}_o^{1/4}, & D_{b1} &= D_o \end{aligned}$$

and

$$\begin{aligned} A_{b2} \mathcal{M}_i^{1/4} &= \mathcal{M}_i^{1/4} A_i, & B_{b2} &= \mathcal{M}_i^{1/4} B_i, \\ C_{b2} \mathcal{M}_i^{1/4} &= C_i, & D_{b2} &= D_i. \end{aligned}$$

2. All parbalanced realizations of G are unitarily equivalent.

3. If $G \in TLD^{U,Y}$ is continuous on $\partial\mathbb{D}$ with values in the set of compact operators, then there exists a balanced realization whose state space is equal to the closure of the range of the Hankel operator with symbol G^\perp . The gramian has a matrix representation with respect to a basis such that its diagonal entries are the singular values of the Hankel operator with symbol G^\perp .

Proof. Statements 2 and 3 and the existence of the first parbalanced realization of statement 1 can be found in [13]. The second realization of statement 1 can be obtained by taking the dual of the parbalanced realization of \bar{G} constructed by the method of the first realization. \square

We have the following proposition concerning the transformation from the restricted ($*$ -restricted) shift realization to the parbalanced realization in Theorem 4.1.

PROPOSITION 4.2. In the notation of Theorem 4.1, the operators $\mathcal{W}_o^{1/2}$ and $\mathcal{W}_o^{1/4}$ are bounded positive definite with dense ranges in X_o ; the operators $\mathcal{M}_i^{1/2}$ and $\mathcal{M}_i^{1/4}$ are bounded positive definite with dense ranges in X_i .

Proof. Clearly, \mathcal{W}_o is a bounded positive definite operator on X_o . Similarly, \mathcal{M}_i is a positive definite operator on X_i . Since

$$\mathcal{W}_o^{1/2}(\mathcal{W}_o^{1/2})^* = H_{G^\perp} H_{G^\perp}^*$$

and $H_{G^\perp} H_{G^\perp}^* X_o$ is dense in X_o , $\mathcal{W}_o^{1/2}$ has dense range in X_o . Hence so does $\mathcal{W}_o^{1/4}$. Similarly, $\mathcal{M}_i^{1/2}$ and $\mathcal{M}_i^{1/4}$ also have dense ranges. \square

Combining Theorem 4.1 and Proposition 4.2, we have, in the terminology of [12], that A_{b1} is a quasi-affine transform of A_o and A_i a quasi-affine transform of A_{b2} .

Theorem 4.1 has some by-products that may be of interest in their own right. First, since two parbalanced realizations are unitarily equivalent, their state spaces must be unitarily equivalent.

COROLLARY 4.3. The spaces $\overline{H_{G^\perp}(H_U^2)}$ and $\overline{H_{\bar{G}^\perp}(H_Y^2)}$ are unitarily equivalent, with a unitary transformation given by

$$V = \mathcal{W}_o^{-1/4} H_{G^\perp} \mathcal{M}_i^{-1/4} : \overline{H_{G^\perp}(H_U^2)} \rightarrow \overline{H_{\bar{G}^\perp}(H_Y^2)}.$$

Before stating the second consequence of Theorem 4.1, we quote from [3, p. 248] the following result regarding the closedness of the range of a Hankel operator.

PROPOSITION 4.4. Let $K \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ with U and Y finite-dimensional. Then $H_K(H_U^2(\mathbb{D}))$ is closed in $H_Y^2(\mathbb{D})$ if and only if there are functions $Q \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$, $F \in H_{\mathcal{L}(Y,U)}^\infty(\mathbb{D})$, $P_1 \in H_{\mathcal{L}(Y)}^\infty(\mathbb{D})$, and $P_2 \in H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ such that, for almost all $z \in \partial\mathbb{D}$,

$$K(z) = Q(z)(zF(z))^*, \quad P_1(z)Q(z) + P_2(z)F(z) = I_Y$$

and Q is inner. Note that the last equality means that Q and F are strongly right coprime [3].

We have the following characterization of when the state space transformation in Theorem 4.1 that maps an input normal (respectively, output normal) realization to a (par-) balanced realization is an equivalence transformation.

PROPOSITION 4.5. *Let $G \in \text{TLD}^{U,Y}$. A parbalanced realization is equivalent to an output normal (or input normal) realization if and only if H_{G^\perp} has closed range.*

Proof. If H_{G^\perp} has closed range, then \mathcal{W}_o in Theorem 4.1 has bounded inverse and $\mathcal{W}_o^{-1/4}$ is an equivalence transformation from the restricted shift realization to a parbalanced realization.

Conversely, if T is an equivalence transformation from a parbalanced realization to the restricted shift realization, then it will follow that $TT^*TT^* = \mathcal{W}_o$. Hence \mathcal{W}_o has bounded inverse. Since $\mathcal{W}_o = H_{G^\perp}H_{G^\perp}^*$, H_{G^\perp} must have closed range. Note that H_{G^\perp} has closed range if and only if $H_{\tilde{G}^\perp}$ has closed range. This completes the proof. \square

In fact, the state space isomorphism theorem holds when H_{G^\perp} has closed range.

COROLLARY 4.6. *Let $G \in \text{TLD}^{U,Y}$. Then all reachable and observable realizations of G are equivalent if and only if H_{G^\perp} has closed range.*

Proof. If all reachable and observable realizations of G are equivalent, then, in particular, the output normal and the parbalanced realizations are equivalent. By Proposition 4.5, H_{G^\perp} has closed range.

Conversely, assume that H_{G^\perp} has closed range. Let $(A, B, C, D) \in \text{LD}_X^{U,Y}$ be a reachable and observable realization of G with state space X . We show that it is equivalent to an output normal realization. Then, by Theorem 3.2, this shows that all reachable and observable realizations of G are equivalent.

Let \mathcal{O} and \mathcal{R} be, respectively, the observability and reachability operators of (A, B, C, D) . It is easily verified that $H_{G^\perp} = \mathcal{O}\mathcal{R} : H_V^2(\mathbb{D}) \rightarrow H_Y^2(\mathbb{D})$ (see the beginning of § 2). Hence $\mathcal{O}\mathcal{R}(H_V^2(\mathbb{D}))$ is closed in $H_Y^2(\mathbb{D})$. By reachability, $\mathcal{R}(H_V^2(\mathbb{D})) \subseteq X$ is dense in X . Thus

$$\mathcal{O}(X) \subseteq \overline{\mathcal{O}\mathcal{R}(H_V^2(\mathbb{D}))} = \mathcal{O}\mathcal{R}(H_V^2(\mathbb{D})) \subseteq \mathcal{O}(X).$$

It follows that $\mathcal{O}(X) = \mathcal{O}\mathcal{R}(H_V^2(\mathbb{D}))$, and hence $\mathcal{O}(X)$ is closed in $H_Y^2(\mathbb{D})$. Since by observability \mathcal{O} is injective, the operator $\mathcal{O} : X \rightarrow \mathcal{O}(X)$ has bounded inverse. Consequently, the operator $\mathcal{O}^*\mathcal{O} : X \rightarrow X$ has bounded inverse on X . Now let $V = (\mathcal{O}^*\mathcal{O})^{-1/2}$, then V is bounded and is boundedly invertible. It is routine to verify that the realization $(V^{-1}AV, V^{-1}B, CV, D)$, which is equivalent to (A, B, C, D) , is output normal. \square

The main result in this section is that all parbalanced realizations are asymptotically stable. We need two lemmas in the proof.

LEMMA 4.7 (see [3, p. 124]). *Let $A : H_1 \rightarrow H$ and $B : H_2 \rightarrow H$ be two linear operators from Hilbert spaces H_1 and H_2 , respectively, into a Hilbert space H . Then $AA^* \leq BB^*$ if and only if there exists a contraction $V : H_1 \rightarrow H_2$ such that $A = BV$. Moreover, $AA^* = BB^*$ if and only if V is a partial isometry with final space equal to $\text{range}(B^*)$.*

LEMMA 4.8. *Let $G \in \text{TLD}^{U,Y}$. Let (A, B, C, D) be a realization of G and let (A^*, C^*, B^*, D^*) be its dual system. Then (A, B, C, D) is a parbalanced realization of G if and only if (A^*, C^*, B^*, D^*) is a parbalanced realization of \tilde{G} .*

THEOREM 4.9. *Let $G \in \text{TLD}^{U,Y}$ and let (A_b, B_b, C_b, D_b) be a parbalanced realization of G . Then $A_b \in C_{00}$.*

Proof. Here we use the notation and result of Theorem 4.1 and first prove that A_{b1} is asymptotically stable. Note that $A_o^* = P_{X_o} S|_{X_o}$ and $X_o^\perp \subseteq \ker(H_{G^\perp}^*)$. It is easy to verify that

$$H_{G^\perp}^* A_o^* = H_{G^\perp}^* S|_{X_o} = S^*|_{X_o} H_{G^\perp}^*|_{X_o} = A_o H_{G^\perp}^*|_{X_o}.$$

Hence we have

$$\begin{aligned} \langle A_o \mathcal{W}_o A_o^* x, x \rangle &= \langle A_o H_{G^\perp} H_{G^\perp}^* A_o^* x, x \rangle \\ &= \langle H_{G^\perp}^* A_o^* x, H_{G^\perp}^* A_o^* x \rangle \\ &= \langle A_o H_{G^\perp}^* x, A_o H_{G^\perp}^* x \rangle \\ &\leq \langle H_{G^\perp}^* x, H_{G^\perp}^* x \rangle \\ &= \langle \mathcal{W}_o x, x \rangle, \end{aligned}$$

i.e., $A_o \mathcal{W}_o A_o^* \leq \mathcal{W}_o$. Thus by Lemma 4.7 there exists a contraction V on X_o such that

$$A_o \mathcal{W}_o^{1/2} = \mathcal{W}_o^{1/2} V,$$

and hence for any positive integer n

$$A_o^n \mathcal{W}_o^{1/2} = \mathcal{W}_o^{1/2} V^n.$$

Let x be any element in $\mathcal{W}_o^{1/2} X_o$, i.e., $x = \mathcal{W}_o^{1/2} z$ for some $z \in X$. Then the element $y = \mathcal{W}_o^{1/4} z \in X_o$ is such that $y = \mathcal{W}_o^{-1/4} x$. The above equality applied to y yields

$$A_o^n \mathcal{W}_o^{1/4} x = \mathcal{W}_o^{1/2} V^n y.$$

Since the right-hand side of the last equality is in $\mathcal{W}_o^{1/2} X_o$, the operator $\mathcal{W}_o^{-1/4}$ can be applied to both sides to lead to

$$\mathcal{W}_o^{-1/4} A_o^n \mathcal{W}_o^{1/4} x = \mathcal{W}_o^{1/4} V^n y.$$

Now, noting that $\mathcal{W}_o^{1/4}$ is selfadjoint and, from Theorem 4.1, $A_{b1}^n x = \mathcal{W}_o^{-1/4} A_o^n \mathcal{W}_o^{1/4} x$, we have that

$$\begin{aligned} \|A_{b1}^n x\|^2 &= \langle A_{b1}^n x, A_{b1}^n x \rangle \\ &= \langle \mathcal{W}_o^{-1/4} A_o^n \mathcal{W}_o^{1/4} x, \mathcal{W}_o^{-1/4} A_o^n \mathcal{W}_o^{1/4} x \rangle \\ &= \langle \mathcal{W}_o^{-1/4} A_o^n \mathcal{W}_o^{1/4} x, \mathcal{W}_o^{1/4} V^n y \rangle \\ &= \langle \mathcal{W}_o^{1/4} \mathcal{W}_o^{-1/4} A_o^n \mathcal{W}_o^{1/4} x, V^n y \rangle \\ &= \langle A_o^n \mathcal{W}_o^{1/4} x, V^n y \rangle \\ &\rightarrow 0 \end{aligned}$$

as $A_o^n z \rightarrow 0$ for any z and $\|V^n y\| \leq \|y\|$.

We thus have proved $\|A_{b1}^n x\| \rightarrow 0$ for any $x \in \mathcal{W}_o^{1/2} X_o$. Let $z \in X_o$ and $\epsilon > 0$. Since $\mathcal{W}_o^{1/2} X_o$ is dense in X_o , there exists $x \in \mathcal{W}_o^{1/2} X_o$ such that $\|z - x\| < \epsilon/2$. Choosing N such that $\|A_{b1}^n x\| < \epsilon/2$ whenever $n \geq N$ and using the fact that A_{b1} is a contraction, we obtain, for $n \geq N$

$$\|A_{b1}^n z\| \leq \|A_{b1}^n (z - x)\| + \|A_{b1}^n x\| < \|A_{b1}^n\| \|z - x\| + \epsilon/2 \leq \epsilon/2 + \epsilon/2 = \epsilon.$$

This shows that A_{b1} is asymptotically stable. Since, by statement 2 of Theorem 4.1, all observable and reachable parbalanced realizations are unitarily equivalent, A_b must be asymptotically stable. Hence any parbalanced realization of any transfer function in $TLD^{U,Y}$ is asymptotically stable. Since by Lemma 4.8 $(A_b^*, C_b^*, B_b^*, D_b^*)$ is a parbalanced realization of the transfer function $\tilde{G} \in TLD^{Y,U}$, we can apply this statement to $(A_b^*, C_b^*, B_b^*, D_b^*)$ to get the asymptotic stability of A_b^* . Therefore we have proved that $A_b \in C_{00}$. \square

We discuss the spectral properties of a parbalanced realization and relate these properties to the characterization of power-stability of parbalanced realizations.

PROPOSITION 4.10. *Let (A_b, B_b, C_b, D_b) , (A_i, B_i, C_i, D_i) , and (A_o, B_o, C_o, D_o) be, respectively, a parbalanced, an input normal and an output normal realization of $G \in TLD^{U,Y}$ with U and Y finite-dimensional. If G^\perp is in $H_{\mathcal{L}(U,Y)}^\infty$ and is strictly noncyclic, then A_b, A_i, A_o are all C_0 operators. Moreover, they have the same minimal function.*

Proof. By Theorem 3.5, Corollary 3.6, and Theorem 4.9, the assumption in the proposition implies that A_i, A_o , and A_b are all in C_{00} . Hence they are all completely nonunitary (see [12] or [8]). Furthermore, as noted after Proposition 4.2, A_{b1} is a quasi-affine transform of A_o and A_i a quasi-affine transform of A_{b2} . The result now follows from Proposition 3.13 and [12, Prop. 4.6, p. 125], which shows the following: For two completely nonunitary operators A and B on a Hilbert space H , if there is a bounded injective operator C on H with dense range in H such that $AC = CB$ (i.e., B is a quasi-affine transform of A), then A is a C_0 operator if and only if B is, and in this case they both have the same minimal function. \square

For the spectrum of the state propagation operators, we obtain the following result.

COROLLARY 4.11. *Under the assumption of Proposition 4.10, we have*

$$\sigma(A_b) = \sigma(A_i) = \sigma(A_o) \quad \text{and} \quad \sigma_p(A_b) = \sigma_p(A_i) = \sigma_p(A_o).$$

Proof. The proof is an immediate consequence of Propositions 4.10 and 3.8. \square

For the question of the spectral minimality, we have the same result as for input normal and output normal realizations in the case of finite-dimensional U and Y .

COROLLARY 4.12. *Under the assumption of Proposition 4.10, the systems (A_b, B_b, C_b, D_b) , (A_i, B_i, C_i, D_i) , and (A_o, B_o, C_o, D_o) are spectrally minimal, i.e.,*

$$\sigma_s(G) = \sigma(A_b) = \sigma(A_i) = \sigma(A_o).$$

Proof. Combining Theorems 3.14 with 3.2, we have that

$$\sigma_s(G) = \sigma(A_i) = \sigma(A_o).$$

Corollary 4.11 now implies the result. \square

The criteria for power-stability are also identical to those in the input normal and output normal case if G^\perp is strictly noncyclic.

COROLLARY 4.13. *Let (A_b, B_b, C_b, D_b) be a parbalanced realization of $G \in TLD^{U,Y}$ with U and Y finite-dimensional. Assume that G^\perp is in $H_{\mathcal{L}(U,Y)}^\infty(\mathbb{D})$ and is strictly noncyclic. Then A_b is power-stable if and only if G is rational.*

Proof. The proof follows from Corollary 4.11 and Theorem 3.15. \square

This corollary shows that a parbalanced realization of $G \in TLD^{U,Y}$, with G^\perp nonrational and strictly noncyclic, cannot be power-stable. When G^\perp is not strictly

noncyclic the situation is complicated. Here we give an example of a power-stable parbalanced realization of a cyclic function with l_2 as its state space.

Example. Let S and S^* be the right and left shifts on the space l_2 . Let $A = \frac{1}{5}(I + S + S^*)$. Clearly, $\|A\| \leq \frac{3}{5}$. Define $B : \mathbb{C} \rightarrow l_2$ as

$$B(\lambda) = (\lambda, 0, 0, \dots)^T, \quad \lambda \in \mathbb{C}$$

and $C : l_2 \rightarrow \mathbb{C}$ as

$$C(x_k)_{k \geq 1} = x_1, \quad (x_k)_{k \geq 1} \in l_2.$$

We take D to be zero. We have $\|B\| = 1$ and $\|C\| = 1$. Now consider $e_i = (\delta_{ij})_{j \geq 1}$, where δ_{ij} is the Kronecker delta. Then $\{e_i\}_{i \geq 1}$ forms a basis of l_2 . With respect to this basis, we have the following matrix representations of A , B , and C :

$$A = \frac{1}{5} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 1 & 0 & 0 & \cdots \\ 0 & 1 & 1 & 1 & 0 & \cdots \\ 0 & 0 & 1 & 1 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}, \quad \text{and} \quad C = [1 \ 0 \ 0 \ 0 \ \cdots].$$

We show that (A, B, C, D) is an observable and reachable system. Let \mathcal{O} and \mathcal{R} be, respectively, the observability and reachability operators. For $x = (x_k)_{k \geq 0} \in l^2$, we have

$$\|\mathcal{O}x\|^2 = \|(CA^k x)_{k \geq 0}\|^2 = \sum_{k \geq 0} \|CA^k x\|^2 \leq \sum_{k \geq 0} \|A\|^{2k} \|x\|^2 \leq \sum_{k \geq 0} \left(\frac{3}{5}\right)^{2k} \|x\|^2.$$

Hence \mathcal{O} is bounded. Let $x = (x_k)_{k \geq 0} \in l^2$ be such that $\mathcal{O}x = 0$, i.e., $CA^k x = 0$ for $k = 0, 1, \dots$. Then it follows that $x_1 = CA^0 x = 0$, and hence $x_2 = 0$ because $0 = CAx = (x_1 + x_2)/5$, and so on. So we have $x = 0$. This shows that the system is observable. Note that $\mathcal{R} = \mathcal{O}^*$. Hence the system is reachable. It is obviously parbalanced. Also, the transfer function $g(z) = C(zI - A)^{-1}B$ is such that $g^\perp \in H^\infty$ due to the fact that $\|A\| < 1$. Since this is a power-stable realization, by Corollary 4.13, g must be cyclic. Thus there exists a cyclic transfer function that has power-stable parbalanced realizations.

5. Concluding remarks. We have shown the asymptotic stability of parbalanced realizations and have given conditions for an input normal or output normal realization to be asymptotically stable. An input normal or output normal realization cannot be power-stable unless the transfer function is rational. This is also true for parbalanced realizations when the transfer functions are assumed to be strictly noncyclic. If the transfer function is cyclic, the problem of finding a full characterization for power stability of parbalanced realizations remains open.

Concluding the paper, we point out that the results here can be translated to continuous-time systems by the bilinear mapping defined in [10]. However, to use that mapping, we restrict the discrete-time transfer functions to be *admissible*. A function G is said to be an admissible discrete-time transfer function if G is in $TL D^{U,Y}$ and the limit

$$\lim_{\substack{\lambda < -1, \lambda \rightarrow -1 \\ \lambda \in \mathbb{R}}} G(\lambda)$$

exists in the norm topology. Correspondingly, the discrete-time linear systems (A, B, C, D) must be *admissible*, also; that is, in addition to A being contractive, B, C , and D being bounded, the limit

$$\lim_{\substack{\lambda \rightarrow -1, \lambda > -1 \\ \lambda \in \mathbb{R}}} C(\lambda I + A)^{-1} B$$

must exist in the norm topology and $-1 \notin \sigma_p(A)$. It can be easily verified that the restricted and *-restricted shift realizations of admissible transfer functions are admissible systems. Moreover, the dual system of an admissible system is admissible, and any reachable and observable parbalanced realization of an admissible transfer function is an admissible system. Since the class of admissible transfer functions (linear systems) is smaller than the class of transfer functions (linear systems) considered in this paper, all the results of this paper are also valid for the smaller class.

REFERENCES

- [1] R. G. DOUGLAS, H. S. SHAPIRO, AND A. L. SHIELDS, *Cyclic vectors and invariant subspaces for the backward shift operator*, Ann. Inst. Fourier (Grenoble), 20 (1970), pp. 37–76.
- [2] P. A. FUHRMANN, *Realization theory in Hilbert space for a class of transfer functions*, J. Funct. Anal., 18 (1975), pp. 338–349.
- [3] ———, *Linear systems and operators in Hilbert space*, McGraw-Hill Inc., 1981.
- [4] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.
- [5] K. GLOVER, R. F. CURTAIN, AND J. R. PARTINGTON, *Realisation and approximation of linear infinite dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.
- [6] J. W. HELTON, *Systems with infinite dimensional state space*, Proc. IEEE, 64 (1976), pp. 145–160.
- [7] B. C. MOORE, *Principal component analysis in linear systems: Controllability, observability and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.
- [8] N. K. NIKOL'SKIĬ, *Treatise on the Shift Operator: Spectral Function Theory*, Springer-Verlag, Berlin, New York, 1986.
- [9] R. OBER, *Balanced parametrizations of classes of linear systems*, SIAM J. Control Optim., 29 (1991), pp. 1251–1287.
- [10] R. OBER AND S. MONTGOMERY-SMITH, *Bilinear transformation of infinite dimensional state space systems and balanced realizations of nonrational transfer functions*, SIAM J. Control Optim., 28 (1990), pp. 439–465.
- [11] K. M. PRZYLUŚKI, *Stability of linear infinite-dimensional systems revisited*, Internat. J. Control, 48 (1988), pp. 513–523.
- [12] B. SZ.-NAGY AND C. FOIAS, *Harmonic analysis of operators on Hilbert space*, North-Holland, Amsterdam, 1970.
- [13] N. YOUNG, *Balanced realizations in infinite dimensions*, in Operator Theory, Advances and Applications 19, Birkhäuser Verlag, Basel, Boston, 1986, pp. 449–470.

WEAK SHARP MINIMA IN MATHEMATICAL PROGRAMMING*

J. V. BURKE[†] AND M. C. FERRIS[‡]

Abstract. The notion of a *sharp*, or *strongly unique*, minimum is extended to include the possibility of a nonunique solution set. These minima will be called *weak sharp minima*. Conditions necessary for the solution set of a minimization problem to be a set of weak sharp minima are developed in both the unconstrained and constrained cases. These conditions are also shown to be sufficient under the appropriate convexity hypotheses. The existence of weak sharp minima is characterized in the cases of linear and quadratic convex programming and for the linear complementarity problem. In particular, a result of Mangasarian and Meyer is reproduced that shows that the solution set of a linear program is always a set of weak sharp minima whenever it is nonempty. Consequences for the convergence theory of algorithms are also examined, especially conditions yielding finite termination.

Key words. finite termination, strongly unique minima, sharp minima

AMS subject classifications. 90C20, 90C30, 65K05

1. Introduction. Let $f: X \mapsto \bar{\mathbf{R}} := \mathbf{R} \cup \{-\infty, \infty\}$; we say that f has a sharp minimum at $\bar{x} \in \mathbf{R}^n$ if $f(x) \geq f(\bar{x}) + \alpha \|x - \bar{x}\|$, for all x near \bar{x} and some $\alpha > 0$. The notion of a sharp minimum, or equivalently, a strongly unique local minimum, has far reaching consequences for the convergence analysis of many iterative procedures [1], [8], [11], [12], [17], [18]. In this article, we extend the notion of a sharp minimum to include the possibility of a nonunique solution set. We say that $\bar{S} \subset \mathbf{R}^n$ is a set of *weak sharp minima* for the function f relative to the set $S \subset \mathbf{R}^n$ where $\bar{S} \subset S$ if there is an $\alpha > 0$ such that

$$(1) \quad f(x) \geq f(y) + \alpha \text{dist}(x | \bar{S}),$$

for all $x \in S$ and $y \in \bar{S}$ where

$$\text{dist}(x | \bar{S}) := \inf_{z \in \bar{S}} \|x - z\|.$$

The constant α and the set \bar{S} are called the modulus and domain of sharpness for f over S , respectively. Clearly, \bar{S} is a set of global minima for f over S . The notion of weak sharp minima is easily localized. We will say that $\bar{x} \in \mathbf{R}^n$ is a local weak sharp minimum for f on $S \subset \mathbf{R}^n$ if there exists a set $\bar{S} \subset S$ and a parameter $\delta > 0$ with $\bar{x} \in \bar{S}$ such that the set $\bar{S} \cap \{x : \|x - \bar{x}\| \leq \delta\}$ is a set of weak sharp minima for the function

$$f_\delta(x) := \begin{cases} f(x), & \text{if } \|x - \bar{x}\| \leq \delta, \\ +\infty, & \text{otherwise,} \end{cases}$$

relative to the set S . Since the restriction to the local setting is straightforward, we will concentrate on the global definition.

The study of weak sharp minima is motivated primarily by applications in convex and convex composite programming, where such minima commonly occur. For example, such minima frequently occur in linear programming, linear complementarity,

* Received by the editors September 30, 1991; accepted for publication (in revised form) April 5, 1992. This material is based on research supported by National Science Foundation grants CCR-9157632 and DMS-9102059 and Air Force Office of Scientific Research grant AFOSR-89-0410.

[†] Department of Mathematics, GN-50, University of Washington, Seattle, Washington 98195.

[‡] Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706.

and least distance or projection problems. The goals of this study are to quantify this property, investigate its geometric structure, characterize its occurrence in simple convex programming problems, and, finally, to analyze its impact on the convergence of algorithms. Furthermore, although our primary interest is with convex programming, we also investigate the significance of weak sharp minima for nonconvex problems. However, in the latter case, rather strong regularity conditions are required to yield significant extensions of the convex case. Nonetheless, we do obtain some very interesting and significant results for differentiable problems with convex constraints. These results extend and refine earlier work of Al-Khayyal and Kyparisis [1] on the finite termination of algorithms at sharp minima. In a later study, we also show how these results can be applied to convex composite optimization problems to establish the quadratic rate of convergence of a variety of algorithms. This study builds on the work initiated in [9].

Our study begins in §2 with the derivation of first-order necessary conditions for the solution set of a problem to be a set of weak sharp minima. The unconstrained ($S = \mathbb{R}^n$) and constrained cases are treated separately. When the problem data is convex, it is shown that these conditions are also sufficient. In the third section these results are applied to three important classes of convex programs: quadratic programming, linear programming, and the linear complementarity problem. In the final section we examine certain tools for studying the convergence of algorithms in the presence of weak sharp minima. In particular, it is shown how we can attain finite convergence to weak sharp minima.

The notation that we employ is for the most part standard; however, a partial list is provided for the readers' convenience. The *inner product* on \mathbb{R}^n is defined as the bilinear form

$$\langle y, x \rangle := \sum_{i=1}^n y_i x_i.$$

We denote a *norm* on \mathbb{R}^n by $\|\cdot\|$. Each norm defines a norm dual to it and is given by

$$\|x\|_o := \sup_{\|y\| \leq 1} \langle y, x \rangle.$$

The associated closed unit balls for these norms are denoted by B and B° , respectively. The 2-norm plays a special role in our development and is denoted by

$$\|x\|_2 := \sqrt{\langle x, x \rangle}.$$

If it is understood from the context that we are speaking of the 2-norm, then we will drop the subscript "2" from this notation.

Given two subsets A and B of \mathbb{R}^n and $\beta \in \mathbb{R}$, we define

$$A \pm \beta B := \{a \pm \beta b : a \in A, b \in B\}.$$

On the other hand,

$$A \setminus B := \{a \in A : a \notin B\}.$$

If $A \subset \mathbb{R}^n$ then the *polar* of A is defined to be the set

$$A^\circ := \{x^* \in \mathbb{R}^n : \langle x^*, x \rangle \leq 1 \ \forall x \in A\}.$$

This notation is consistent with the definition of the dual unit ball B° . The *indicator* and *support* functions for A are given by

$$\psi(x \mid A) := \begin{cases} 0 & \text{if } x \in A, \\ +\infty & \text{otherwise} \end{cases}$$

and

$$\psi^*(x \mid A) := \sup\{\langle x^*, x \rangle : x^* \in A\},$$

respectively. Moreover, we write $\text{int } A$ for the interior of A , $\text{cl } A$ for the closure of A , and $\text{span } A$ for the linear span of the elements of A . The *relative interior* of A , denoted $\text{ri } A$, is the interior of A relative to the *affine hull* of A , which is given by

$$\text{aff } A := \left\{ \sum_{k=1}^s \lambda_k x_k \mid \begin{array}{l} s \in \{1, 2, \dots\}, x_k \in A \text{ and } \lambda_k \in \mathbb{R} \\ \text{for } k = 1, 2, \dots, s, \text{ with } \sum_{k=1}^s \lambda_k = 1 \end{array} \right\}.$$

The subspace *perpendicular* to A is defined to be

$$A^\perp := \{y \in \mathbb{R}^n : \langle y, x \rangle = 0 \text{ for all } x \in A\}.$$

If A is closed, then we define the *projection* of a point $x \in \mathbb{R}^n$ onto the set A as the set of all points in A that are closest to x in a given norm. In this paper, we will only speak of the projection with respect to the 2-norm; it is denoted by

$$P(x \mid A) := \{\bar{y} \in A : \|x - \bar{y}\|_2 = \inf_{y \in A} \|x - y\|_2\}.$$

The projection is an example of a multivalued mapping on \mathbb{R}^n . The set A is said to be *convex* if the line segment connecting any two points in A is also contained in A . The *convex hull* of the set A , denoted $\text{co}(A)$, is the smallest convex set that contains A ; that is, $\text{co}(A)$ is the intersection of all convex sets that contain A . It is interesting to note that the projection operator can be used to characterize the closed convex subsets of \mathbb{R}^n . That is, the set A is closed and convex if and only if the projection operator for A , $P(\cdot \mid A)$, is single valued on all of \mathbb{R}^n [2], [16].

Given $x \in A$, we define the *normal cone* to A at x , denoted $N(x \mid A)$, to be the closure of the convex hull of all limits of the form

$$\lim_k t_k^{-1}(x_k - p_k),$$

where the sequences $\{t_k\} \subset \mathbb{R}$, $\{p_k\} \subset A$, and $\{x_k\} \subset \mathbb{R}^n$ satisfy $t_k \downarrow 0$, $p_k \in P(x_k \mid A)$, and $p_k \rightarrow x$. If A is convex, we can show that this definition implies that

$$N(x \mid A) = \{x^* \in \mathbb{R}^n : \langle x^*, y - x \rangle \leq 0 \ \forall y \in A\}.$$

The *tangent cone* to A at x is defined dually by the relation

$$T(x \mid A) := N(x \mid A)^\circ.$$

If A is convex, we have the relation

$$T(x \mid A) = \text{cl} [\cup_{\lambda \geq 0} \lambda(A - x)].$$

The *contingent cone* to A at x plays a role similar to that of the tangent cone but is, in general, larger. The contingent cone to A at x is given by

$$K(x | A) := \{d \in \mathbb{R}^n : \exists t_k \downarrow 0, d^k \rightarrow d, \text{ with } x + t_k d^k \in A\}.$$

The set A is said to be *regular* at $x \in A$ if $T(x | A) = K(x | A)$. In particular, every convex set is regular.

Let $f: X \mapsto \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. The *domain* and *epigraph* of f are given by

$$\text{dom } f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$$

and

$$\text{epi } f := \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \lambda\},$$

respectively. Observe that f is lower semicontinuous if and only if $\text{epi } f$ is closed. For $x \in \text{dom } f$, we define the *subdifferential* of f at x to be the set

$$\partial f(x) := \{x^* : (x^*, -1) \in N((x, f(x)) | \text{epi } f)\},$$

and the *singular subdifferential* of f at x to be the set

$$\partial^\infty f(x) := \{x^* : (x^*, 0) \in N((x, f(x)) | \text{epi } f)\}.$$

The mappings ∂f and $\partial^\infty f$ are further examples of multivalued mappings on \mathbb{R}^n . We observe that the set $\partial f(x) \cup \partial^\infty f(x)$ is always nonempty even though ∂f may be empty at certain points. Moreover, the function f is locally Lipschitzian on \mathbb{R}^n if and only if ∂f is nonempty and compact valued on all of \mathbb{R}^n . The domain of ∂f is the set

$$\text{dom } \partial f := \{x^* \in \mathbb{R}^n : \partial f(x) \neq \emptyset\}.$$

If f is convex, then this subdifferential coincides with the usual subdifferential from convex analysis. The *generalized directional derivative* of f is the support function of $\partial f(x)$,

$$f^\circ(x; d) := \psi^*(d | \partial f(x)),$$

and the *contingent directional derivative* of f at x in the direction d is given by

$$f^-(x; d) := \liminf_{\substack{u \rightarrow d \\ t \downarrow 0}} \frac{f(x + tu) - f(x)}{t}.$$

The relation $f^-(x; d) \leq f^\circ(x; d)$ always holds. The function f is said to be *regular* at x if $f^\circ(x; d) = f^-(x; d)$ in which case the usual directional derivative,

$$f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t},$$

exists and equals this common value. See [7] for further details of subdifferential calculus.

2. Subdifferential geometry. We begin with a study of some geometric consequences of weak sharp minima. Specifically, we are interested in first-order necessary conditions. The general unconstrained ($S = \mathbb{R}^n$) and constrained cases are treated separately. In both cases, it is shown that the necessary conditions are also sufficient under appropriate convexity hypotheses. The following preliminary result is required.

LEMMA 2.1. *Suppose $f: \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ is closed, proper, and convex, the sets $\bar{S} := \arg \min \{f(x) : x \in \mathbb{R}^n\}$ and C are nonempty, closed, and convex subsets of \mathbb{R}^n with $C \subseteq \bar{S}$, and $\alpha > 0$. The following are equivalent:*

1. $\alpha B^\circ \cap N(x \mid \bar{S}) \subseteq \partial f(x)$, for all $x \in C$,
2. $\alpha B^\circ \cap \bigcup_{x \in C} N(x \mid \bar{S}) \subseteq \bigcup_{x \in C} \partial f(x)$.

Proof. [1 \implies 2]. Trivial.

[2 \implies 1]. Let $z \in C$ and $z^* \in \alpha B^\circ \cap N(z \mid \bar{S})$. Then by hypothesis, $z^* \in \partial f(u)$ for some $u \in C$. Since $C \subseteq \bar{S}$ implies that $\partial f(u) \subseteq N(u \mid \bar{S})$, hence $z^* \in N(u \mid \bar{S})$, it follows from $z^* \in N(z \mid \bar{S})$ that

$$(2) \quad \langle z^*, u \rangle = \langle z^*, z \rangle.$$

However, $z^* \in \partial f(u)$ is by definition $f(y) - f(u) \geq \langle z^*, y - u \rangle$, for all y . Since $u, z \in \bar{S}$, $f(u) = f(z)$ so that (2) gives $f(y) - f(z) \geq \langle z^*, y - z \rangle$, for all y , or equivalently, $z^* \in \partial f(z)$. \square

Necessary conditions for weak sharp minima in the unconstrained case now follow.

THEOREM 2.2. *Let $f: \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ be lower semicontinuous and $\alpha > 0$. Consider the following statements:*

1. *The set \bar{S} is a set of weak sharp minima for the function f on \mathbb{R}^n with modulus α .*
2. *For all $d \in \mathbb{R}^n$,*

$$f^-(x; d) \geq \alpha \text{dist}(d \mid K(x \mid \bar{S})).$$

3. *For all $d \in \mathbb{R}^n$*

$$f^\circ(x; d) \geq \alpha \text{dist}(d \mid T(x \mid \bar{S})).$$

4. *The inclusion*

$$\alpha B^\circ \cap N(x \mid \bar{S}) \subseteq \partial f(x)$$

holds.

5. *The inclusion*

$$\alpha B^\circ \cap \left[\bigcup_{x \in \bar{S}} N(x \mid \bar{S}) \right] \subseteq \bigcup_{x \in \bar{S}} \partial f(x)$$

holds.

6. *For all $y \in \mathbb{R}^n$,*

$$f'(p; y - p) \geq \alpha \text{dist}(y \mid \bar{S}),$$

where $p \in P(y \mid \bar{S})$.

Statement 1 implies statement 2 for all $x \in \bar{S}$. Statement 2 implies statement 3 at points $x \in \bar{S}$ at which \bar{S} is regular. Statements 3 and 4 are equivalent. If f is closed proper and convex and the set \bar{S} is nonempty closed and convex, then statements 1–6 are equivalent with 2, 3, and 4 holding at every point of \bar{S} .

Proof.

[1 \implies 2]. Let $x \in \bar{S}$. The hypothesis guarantees that for all t and d'

$$f(x + td') - f(x) \geq \alpha \text{dist}(x + td' \mid \bar{S}),$$

which implies that

$$\frac{f(x + td') - f(x)}{t} \geq \alpha \frac{\text{dist}(x + td' \mid \bar{S}) - \text{dist}(x \mid \bar{S})}{t}.$$

By taking lim infs of both sides as $d' \rightarrow d$ and $t \downarrow 0$ and applying [4, Thm. 4], we obtain the result.

[(2 plus regularity) \implies 3]. Simply observe that regularity at $x \in \bar{S}$ implies the equivalence $T(x \mid \bar{S}) = K(x \mid \bar{S})$ and by definition $f^\circ(x; \cdot) \geq f^-(x; \cdot)$.

[3 \iff 4]. We recall from [5, Thm. 3.1] that if $K \subset \mathbb{R}^n$ is a nonempty closed convex cone, then

$$\text{dist}(x \mid K) = \psi^*(x \mid K^\circ \cap B^\circ).$$

The result now follows from the fact that $f^\circ(x; \cdot) = \psi^*(\cdot \mid \partial f(x))$.

Observe that if f is closed proper and convex, and \bar{S} is nonempty closed and convex, then f is regular on its domain and \bar{S} is regular at each of its elements. Hence either one of the statements 1 or 2 implies both 3 and 4 for all $x \in \bar{S}$.

[(4 holds for all $x \in \bar{S}$) \implies 5]. Trivial.

[(5 plus convexity) \implies 4]. Convexity and Lemma 2.1 combine to establish that 5 implies 4.

[(5 plus convexity) \implies 1]. Given $y \in \mathbb{R}^n$, Theorem 1 in [4] implies the existence of a $x^* \in \alpha B^\circ \cap N(P(y \mid \bar{S}) \mid \bar{S})$ such that $\alpha \text{dist}(y \mid \bar{S}) = \langle x^*, y \rangle - \psi^*(x^* \mid \bar{S})$. Thus, by hypothesis, there exists a $x \in \bar{S}$ with $x^* \in \partial f(x)$. Hence

$$\begin{aligned} f(y) &\geq f(x) + \langle x^*, y - x \rangle \\ &\geq f(x) + \langle x^*, y \rangle - \langle x^*, x \rangle \\ &\geq f(x) + \langle x^*, y \rangle - \psi^*(x^* \mid \bar{S}) \\ &= f(x) + \alpha \text{dist}(y \mid \bar{S}). \end{aligned}$$

Since $y \in \mathbb{R}^n$ is arbitrary, the result is obtained.

[(1 plus convexity) \implies 6]. Let y be given and define $p := P(y \mid \bar{S})$ so that $f(y) \geq f(p) + \alpha \text{dist}(y \mid \bar{S}) = f(p) + \alpha \|y - p\|$. Let $z = \lambda y + (1 - \lambda)p$ for $\lambda \in [0, 1]$. Then $p = P(z \mid \bar{S})$ and

$$f(z) \geq f(p) + \alpha \|z - p\| = f(p) + \alpha \lambda \|y - p\|$$

implying that

$$\frac{f(p + \lambda(y - p)) - f(p)}{\lambda} \geq \alpha \|y - p\|.$$

The result now follows in the limit.

[(6 plus convexity) \implies 1]. Since f is convex it follows that for all x and y

$$f'(x; y - x) = \inf_{t>0} \left[\frac{f(x + t(y - x)) - f(x)}{t} \right]$$

so that for any y we may take $x = P(y \mid \bar{S}) = p$, $t = 1$ and

$$f(p + y - p) - f(p) \geq f'(p; y - p) \geq \alpha \text{dist}(y \mid \bar{S}). \quad \square$$

COROLLARY 2.3. *Suppose f is closed proper and convex and has a set of weak sharp minima \bar{S} that is nonempty, closed, convex, and compact. Then*

$$0 \in \text{int} \bigcup_{\bar{x} \in \bar{S}} \partial f(\bar{x}).$$

Proof. The corollary follows if we can show that

$$\bigcup_{x \in \bar{S}} N(x \mid \bar{S}) = \mathbf{R}^n.$$

Clearly, $\bigcup_{x \in \bar{S}} N(x \mid \bar{S}) \subset \mathbf{R}^n$, so let $y \in \mathbf{R}^n$. By continuity of $\langle y, \cdot \rangle$ and compactness of \bar{S}

$$z^* \in \arg \max_{z \in \bar{S}} \langle y, z \rangle$$

so that $\langle y, z - z^* \rangle \leq 0$, for all $z \in \bar{S}$. Hence $y \in N(z^* \mid \bar{S})$. \square

In the constrained case, we must introduce a constraint qualification to guarantee the validity of the type of first-order optimality conditions that are required for our analysis. For the problem

$$(3) \quad \begin{array}{ll} \text{minimize} & f(x), \\ & x \in S \end{array}$$

these optimality conditions take the form

$$(4) \quad 0 \in \partial f(x) + N(x \mid S).$$

Condition (4) is not always guaranteed to be valid even in the fully convex case, so a constraint qualification is required.

Example 2.4. Consider (3), where $f: \mathbf{R} \mapsto \bar{\mathbf{R}}$ is given by

$$f(x) := \begin{cases} -\sqrt{1+x^2}, & \text{for } x \in [-1, 1] \\ +\infty, & \text{otherwise,} \end{cases}$$

and $S := \{x : x \leq -1\}$. This is a convex program with a closed proper convex objective function having unique global solution $\bar{x} = -1$. However, (4) does not hold since $\partial f(\bar{x}) = \emptyset$.

For this reason we introduce the following constraint qualification due to Rockafellar [20].

DEFINITION 2.5. We say that the *basic constraint qualification* (BCQ) for (3) is satisfied at $x \in S$ if for every $u \in \partial^\infty f(x)$ and $v \in N(x \mid S)$ such that $u + v = 0$ it must be the case that $u = v = 0$. The BCQ is said to be satisfied on a set $\bar{S} \subset S$ if it is satisfied at every point of \bar{S} .

From Rockafellar [20, Cor. 5.2.1], we know that the optimality condition (4) is satisfied at every local solution to (3) at which the BCQ holds. In particular, if f is locally Lipschitzian on \mathbb{R}^n , then $\partial^\infty f(x) = \{0\}$ on all of \mathbb{R}^n ; hence the BCQ is vacuously satisfied on all of S , so (4) holds at every local minima for (3).

THEOREM 2.6. *Suppose $f: \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ is lower semicontinuous and $\bar{S} \subset S$ are nonempty closed subsets of \mathbb{R}^n .*

(a) *The inclusion*

$$(5) \quad \alpha B \subset \partial f(\bar{x}) + \left[T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}) \right]^\circ.$$

holds at $\bar{x} \in \bar{S}$ if and only if

$$(6) \quad f^\circ(\bar{x}; z) \geq \alpha \|z\| \quad \forall z \in T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}).$$

(b) *If \bar{S} is a set of weak sharp minima for f over S with modulus $\alpha > 0$ such that the BCQ holds at every point of \bar{S} , then for each $\bar{x} \in \bar{S}$ at which f , S , and \bar{S} are regular, we have the inclusion (5).*

(c) *If we further assume that f is closed proper and convex and the sets \bar{S} and S are nonempty closed and convex, then \bar{S} is a set of weak sharp minima for f over S with modulus $\alpha > 0$ if and only if the inclusion (5) holds for all $\bar{x} \in \bar{S}$.*

Proof. (a) We show that (5) and (6) are equivalent. Clearly, both statements are false if $\partial f(\bar{x})$ is empty, so we assume it to be nonempty. First note that (6) is equivalent to

$$(7) \quad \sup \{ \langle x^*, z \rangle \mid x^* \in \partial f(\bar{x}) \} \geq \alpha \|z\| \quad \forall z \in T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}).$$

We show this is equivalent to

$$(8) \quad \sup \left\{ \langle x^*, z \rangle \mid x^* \in \partial f(\bar{x}) + \left[T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}) \right]^\circ \right\} \geq \alpha \|z\| \quad \forall z \in \mathbb{R}^n.$$

This is accomplished in two parts. First, it is shown that the supremum in (8) is infinite if $z \notin T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S})$ and then it is shown that the suprema in (7) and (8) are equal if $z \in T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S})$. Suppose $z \notin T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S})$. Then there exists $z^* \in \left[T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}) \right]^\circ$ such that $\langle z^*, z \rangle > 0$. Let $x^* \in \partial f(\bar{x})$, which is nonempty by assumption, and consider $x^* + \lambda z^*$ as $\lambda \rightarrow \infty$. Since $\langle x^* + \lambda z^*, z \rangle \uparrow +\infty$ as $\lambda \uparrow +\infty$, we see that the supremum in (8) is infinite. Suppose that $z \in T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S})$. Then

$$\begin{aligned} & \sup \{ \langle x^*, z \rangle \mid x^* \in \partial f(\bar{x}) \} \\ & \leq \sup \left\{ \langle x^*, z \rangle \mid x^* \in \partial f(\bar{x}) + \left[T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}) \right]^\circ \right\} \end{aligned}$$

since $0 \in \left[T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}) \right]^\circ$. However

$$\begin{aligned} & \sup \left\{ \langle x^*, z \rangle \mid x^* \in \partial f(\bar{x}) + \left[T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}) \right]^\circ \right\} \\ & = \sup \left\{ \langle y^* + z^*, z \rangle \mid y^* \in \partial f(\bar{x}), z^* \in \left[T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}) \right]^\circ \right\} \\ & \leq \sup \{ \langle y^*, z \rangle \mid y^* \in \partial f(\bar{x}) \} \end{aligned}$$

by the definition of a polar cone.

Note that (8) is equivalent to

$$\psi^*(z \mid \alpha B) \leq \psi^*(z \mid \partial f(\bar{x}) + [T(\bar{x} \mid S) \cap N(\bar{x} \mid \bar{S})]^\circ),$$

which is equivalent to

$$\alpha B \subseteq \partial f(\bar{x}) + [T(\bar{x} \mid S) \cap N(\bar{x} \mid \bar{S})]^\circ,$$

which establishes the result.

(b) The definitions imply that \bar{S} is a set of weak sharp minima for f over S with modulus $\alpha > 0$ if and only if \bar{S} is a set of weak sharp minima for the function $h(x) := f(x) + \psi(x \mid S)$ over \mathbb{R}^n with modulus $\alpha > 0$. We will show that this implies (6) for every $\bar{x} \in \bar{S}$ at which f , \bar{S} and S are regular.

Let $\bar{x} \in \bar{S}$ be a point at which f , \bar{S} , and S are regular. Since \bar{S} is a set of weak sharp minima for h over \mathbb{R}^n with modulus $\alpha > 0$, Theorem 2.2 implies that

$$h'(\bar{x}; d) \geq \alpha \text{dist}(d \mid T(\bar{x} \mid \bar{S})) \quad \text{for all } d.$$

Now, by the BCQ, [20, Cor. 8.1.2], and the regularity of S , we know that

$$(9) \quad h'(\bar{x}; d) \leq f'(\bar{x}; d) + \psi(\cdot \mid S)'(\bar{x}; d) = f'(\bar{x}; d) + \psi(d \mid T(\bar{x} \mid S)).$$

Therefore,

$$f'(\bar{x}; d) \geq \alpha \text{dist}(d \mid T(\bar{x} \mid \bar{S})) \quad \text{for all } d \in T(\bar{x} \mid S).$$

This last inequality implies (6) since

$$\text{dist}(d \mid T(\bar{x} \mid \bar{S})) = \|d\|$$

for every $d \in N(\bar{x} \mid \bar{S})$.

(c) Since convexity implies regularity, half of this result has already been established in part (b). It remains to show that (5) holding for all $\bar{x} \in \bar{S}$ implies that \bar{S} is a set of weak sharp minima for f over S with modulus α .

Let $\bar{x} \in \bar{S}$. It was shown in part (a) that the statement (5) is equivalent to the statement (6). Thus we need only show that if (6) holds for all $\bar{x} \in \bar{S}$, then \bar{S} is a set of weak sharp minima for f over S with modulus α . To this end, let $x \in \mathbb{R}^n$ be given and set $\bar{x} = P(x \mid \bar{S})$. By (9) we only need consider cases where $x - \bar{x} \in T(\bar{x} \mid S)$. From the definition of projection it follows that $x - \bar{x} \in N(\bar{x} \mid \bar{S})$. Therefore, $f'(\bar{x}; x - \bar{x}) \geq \alpha \|x - \bar{x}\|$, for all x and hence $h'(\bar{x}; x - \bar{x}) \geq \alpha \text{dist}(x \mid \bar{S})$, for all x . By Theorem 2.2, \bar{S} is a set of weak sharp minima for f over S with modulus α . \square

COROLLARY 2.7. *Suppose $f: \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable and $\bar{S} \subset S$ are nonempty closed subsets of \mathbb{R}^n .*

(a) *The inclusion*

$$(10) \quad \alpha B \subset \nabla f(\bar{x}) + [T(\bar{x} \mid S) \cap N(\bar{x} \mid \bar{S})]^\circ$$

holds at $\bar{x} \in \bar{S}$ if and only if

$$\langle \nabla f(\bar{x}), z \rangle \geq \alpha \|z\| \quad \forall z \in T(\bar{x} \mid S) \cap N(\bar{x} \mid \bar{S}).$$

(b) If \bar{S} is a set of weak sharp minima for f over S with modulus $\alpha > 0$, then for each $\bar{x} \in \bar{S}$ at which S and \bar{S} are regular, we have the inclusion (10).

(c) If we further assume that f is closed proper and convex and the sets \bar{S} and S are nonempty closed and convex, then \bar{S} is a set of weak sharp minima for f over S with modulus $\alpha > 0$ if and only if

$$-\nabla f(\bar{x}) \in \text{int} \bigcap_{x \in \bar{S}} [T(x | S) \cap N(x | \bar{S})]^\circ.$$

Remark. The corollary given above is a strengthening of [1, Prop. 2.2]. In particular, the equivalence in part (a) is proven without assumptions on convexity of S . In fact, under the convexity assumptions in part (c), the condition given in [1] is equivalent to strong uniqueness. By relaxing strong uniqueness to the assumption of a weak sharp minimum, all the results of [1, Prop. 2.2] still follow, with the exception of uniqueness.

3. Some special cases. We now examine three important classes of convex programming problems and characterize when these problems possess weak sharp minima. The problem classes considered are linear and quadratic programming and the linear complementarity problem.

3.1. Quadratic programming. We will use the results on weak sharp minima from §2 to obtain a necessary and sufficient condition for weak sharp minima to occur in convex quadratic programs.

The quadratic programming problem is

$$(11) \quad \begin{array}{ll} \text{minimize} & \frac{1}{2} \langle x, Qx \rangle + \langle c, x \rangle, \\ & x \in S \end{array}$$

where S is polyhedral and $Q \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite. The key to our characterization of when problem (11) has weak sharp minima is the relation (6) in Theorem 2.6. To apply this result, we must first obtain a tractable description of the tangent cone to the solution set of (11). This is accomplished by using the description of the solution set of a convex program given in [3], [14].

THEOREM 3.1. *Let \bar{S} be the set of solutions to the problem $\min\{f(x) : x \in S\}$ where both $f: \mathbb{R}^n \mapsto \mathbb{R}$ and $S \subset \mathbb{R}^n$ are taken to be convex and choose $\bar{x} \in \bar{S}$. Then*

$$\bar{S} = \{x \in S \mid \nabla f(x) = \nabla f(\bar{x}), \langle \nabla f(\bar{x}), x - \bar{x} \rangle = 0\}.$$

It is clear that for convex quadratic programs this gives the solution set as

$$\bar{S} = S \cap \{x \mid \langle \nabla f(\bar{x}), x - \bar{x} \rangle = 0\} \cap \{x \mid \nabla^2 f(\bar{x})(x - \bar{x}) = 0\}$$

and since S is polyhedral

$$(12) \quad T(x \mid \bar{S}) = T(x \mid S) \cap (\nabla f(\bar{x}))^\perp \cap \ker(\nabla^2 f(\bar{x})).$$

Note that $\nabla f(\bar{x})$ is constant on the solution set of a convex program and $\nabla^2 f(\bar{x})$ is constant for the problem (11). In the rest of this paper, we will use the notation $\nabla f(\bar{x})$, $\nabla^2 f(\bar{x})$ for these constants and $\text{span}(d)$, $\ker(A)$ to represent the subspace generated by d and the nullspace of the matrix A , respectively.

THEOREM 3.2. Let \bar{S} be the set of solutions to (11) and assume that \bar{S} is non-empty. Then \bar{S} is a set of weak sharp minima for f over S if and only if

$$(\ker(\nabla^2 f(\bar{x})))^\perp \subseteq \text{span}(\nabla f(\bar{x})) + N(x | S), \quad \forall x \in \bar{S},$$

or, equivalently,

$$(\nabla f(\bar{x}))^\perp \cap T(x | S) \subseteq \ker(\nabla^2 f(\bar{x})), \quad \forall x \in \bar{S},$$

where \bar{x} is any element of \bar{S} .

Proof.

(\Leftarrow) We show that (6) holds. Let $x \in \bar{S}$ and $d \in T(x | S)$. Note that (12) and the hypothesis gives

$$\begin{aligned} K &:= T(x | \bar{S})^\circ = N(x | S) + \text{span}(\nabla f(\bar{x})) + (\ker(\nabla^2 f(\bar{x})))^\perp \\ &= N(x | S) + \text{span}(\nabla f(\bar{x})). \end{aligned}$$

Therefore

$$\begin{aligned} \alpha \text{dist}(d | T(x | \bar{S})) &= \alpha \psi^*(d | \mathbf{B} \cap T(x | \bar{S})^\circ) \\ &= \alpha \sup \left\{ \langle z, d \rangle \mid z \in \mathbf{B} \cap K \right\}. \end{aligned}$$

It follows from [21, p. 65] that $K = \text{span}(\nabla f(\bar{x})) + (K \cap (\nabla f(\bar{x}))^\perp)$, hence, $z \in \mathbf{B} \cap K$ implies $z = \lambda \nabla f(\bar{x}) + y$ with $|\lambda| \leq \eta$, where

$$\eta := \begin{cases} 1/\|\nabla f(\bar{x})\|, & \text{if } \|\nabla f(\bar{x})\| \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and $y \in K \cap (\nabla f(\bar{x}))^\perp$. Therefore

$$\begin{aligned} \alpha \text{dist}(d | T(x | \bar{S})) &= \alpha \sup \left\{ \langle \lambda \nabla f(\bar{x}) + y, d \rangle \mid |\lambda| \leq \eta, y \in N(x | S) \cap (\nabla f(\bar{x}))^\perp \right\} \\ &\leq \alpha \eta \langle \nabla f(\bar{x}), d \rangle \\ &\leq \langle \nabla f(\bar{x}), d \rangle = \langle \nabla f(x), d \rangle = f'(x; d) \end{aligned}$$

as required. The last two inequalities follow since d and y are polar to each other and by choosing $\alpha \leq \|\nabla f(\bar{x})\|$ when $\nabla f(\bar{x}) \neq 0$.

(\Rightarrow) Suppose that for some $x \in \bar{S}$, $T(x | S) \cap (\nabla f(\bar{x}))^\perp \not\subseteq \ker(\nabla^2 f(\bar{x}))$. Then there exists $d \in T(x | S) \cap (\nabla f(\bar{x}))^\perp$ with $d \notin \ker(\nabla^2 f(\bar{x}))$. Thus from (12), $d \notin T(x | \bar{S})$ and so

$$\alpha \text{dist}(d | T(x | \bar{S})) > 0 = \langle \nabla f(\bar{x}), d \rangle = f'(x; d),$$

which, using (6), implies that (11) does not have a weak sharp minimum. \square

It is possible to illustrate the theorem by means of adapting a simple example given in [18, p. 206].

Example 3.3. The problem is

$$\begin{aligned} &\text{minimize}_{x \in \mathbb{R}^3} && \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 \\ &\text{subject to} && x_i \in [a_i, b_i], i = 1, 2, 3 \end{aligned}$$

for given $a, b \in \mathbb{R}^3$ with $a \leq b$. We let $S = [a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$. Note that

$$\bar{S} = \{(P(0 \mid [a_1, b_1]), P(0 \mid [a_2, b_2]), x_3) \mid x_3 \in [a_3, b_3]\}$$

and for each $\bar{x} \in \bar{S}$, $\nabla f(\bar{x}) = (\bar{x}_1, \bar{x}_2, 0)$. Also,

$$\nabla^2 f(\bar{x}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ so that } \ker \nabla^2 f(\bar{x}) = \{0\} \times \{0\} \times \mathbb{R}.$$

Furthermore, for $\bar{x} \in S$ we have

$$T(\bar{x} \mid S) = I_1(\bar{x}_1) \times I_2(\bar{x}_2) \times I_3(\bar{x}_3),$$

where

$$I_j(x_j) = \begin{cases} [0, +\infty) & \text{if } x_j = a_j, \\ \mathbb{R} & \text{if } a_j < x_j < b_j, \\ (-\infty, 0] & \text{if } x_j = b_j. \end{cases}$$

It follows that the second equivalence of Theorem 3.2 is satisfied exactly when $0 > b_i$ or $0 < a_i$ for $i = 1, 2$; that is, the box does not straddle the x_1 or x_2 axis. This is precisely when the problem has a weak sharp minimum.

A generalization of this result that does not require the set S to be polyhedral is easily obtained. Observe that the argument given above only employs the polyhedrality of S to establish that (12) holds. However, (12) also holds under the assumption

$$\text{ri}(S - \bar{x}) \cap (\nabla f(\bar{x}))^\perp \cap \ker(\nabla^2 f(\bar{x})) \neq \emptyset$$

(see [21, Cors. 23.8.1 and 16.4.2]), so the following result is immediate.

THEOREM 3.4. *Let \bar{S} be the solution set for (11) where it is no longer assumed that S is polyhedral. Suppose $\bar{x} \in \bar{S}$ is such that*

$$\text{ri}(S - \bar{x}) \cap (\nabla f(\bar{x}))^\perp \cap \ker(\nabla^2 f(\bar{x})) \neq \emptyset.$$

Then \bar{S} is a set of weak sharp minima for f over S if and only if

$$(\ker(\nabla^2 f(\bar{x})))^\perp \subseteq \text{span}(\nabla f(\bar{x})) + N(x \mid S) \quad \forall x \in \bar{S}.$$

3.2. Linear programming. It was shown in [15] that the solution set of a linear program is a set of weak sharp minima. We show below how it can be obtained as a corollary to Theorem 3.2.

The linear programming problem is

$$(13) \quad \begin{array}{ll} \text{minimize} & \langle c, x \rangle \\ & x \in S \end{array}$$

where S is polyhedral.

THEOREM 3.5. *If (13) has a solution, then the set of solutions is a set of weak sharp minima for this problem.*

Proof. Let \bar{x} be a solution of (13). We note that for linear programming $f(x) = \langle c, x \rangle$ so that

$$(\ker(\nabla^2 f(\bar{x})))^\perp = \{0\}.$$

It follows that

$$(\ker(\nabla^2 f(\bar{x})))^\perp \subseteq \text{span}(\nabla f(\bar{x}) + N(x \mid S)) \quad \forall x \in \bar{S},$$

so by Theorem 3.2, (13) has a weak sharp minimum. \square

As was done in Theorem 3.4, we can generalize this result to the case where S is not assumed to be polyhedral.

Remark. It is tempting to consider parametric results for weak sharp minima. In fact, the following example shows that this is not too fruitful. Consider the linear programs $P(i)$, for $i = 1, \dots, \infty$, given by

$$\begin{array}{ll} \text{minimize} & x_1/i + x_2 \\ \text{subject to} & x \geq 0 \end{array}$$

Then, as shown above, each of these problems has a weak sharp minimum. However, it is easy to show that there is no constant $\alpha > 0$ that will work for all of them.

As a simple application of this result, we have the following corollary.

COROLLARY 3.6. *Suppose $f: \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ is a proper polyhedral convex function and the problem*

$$(14) \quad \min_{x \in \mathbb{R}^n} f(x)$$

has a nonempty solution set, \bar{S} . Then \bar{S} is a set of weak sharp minima for (14).

Proof. It follows from the definition of a polyhedral convex function that

$$f(x) = h(x) + \psi(x \mid C),$$

where

$$h(x) := \max\{\langle x, b_1 \rangle - \beta_1, \dots, \langle x, b_k \rangle - \beta_k\}$$

and

$$C := \{x : \langle x, b_{k+1} \rangle \leq \beta_{k+1}, \dots, \langle x, b_m \rangle \leq \beta_m\}.$$

It is clear that (14) is equivalent to

$$(15) \quad \begin{array}{ll} \text{minimize}_x & h(x) \\ \text{subject to} & x \in C, \end{array}$$

which in turn is equivalent to the linear program

$$(16) \quad \begin{array}{ll} \text{minimize}_{(x, \psi)} & \psi \\ \text{subject to} & \begin{array}{l} \psi \geq \langle x, b_i \rangle - \beta_i \quad i = 1, \dots, k \\ x \in C \end{array} \end{array}$$

and that the solution set of (16) is $\bar{S} \times \{h(\bar{x})\}$ for any $\bar{x} \in \bar{S}$. Theorem 3.5 implies the existence of $\alpha > 0$ such that

$$\begin{aligned} \psi - h(\bar{x}) &\geq \alpha \text{dist}((x, \psi) \mid \bar{S} \times \{h(\bar{x})\}) \\ &\geq \alpha \text{dist}(x \mid \bar{S}) \end{aligned}$$

for all (x, ψ) feasible for (16). It then follows that

$$h(x) - h(\bar{x}) \geq \alpha \text{dist}(x \mid \bar{S})$$

for all $x \in C$ since $(x, h(x))$ is feasible for (16). Thus (15) has a weak sharp minimum as required. \square

3.3. Sharpness for linear complementarity problems. We will use the analysis given previously to show that nondegenerate monotone linear complementarity problems have weak sharp minima. This was proved in [13].

The linear complementarity problem is to find an $x \geq 0$ with $Mx + q \geq 0$ satisfying $\langle x, Mx + q \rangle = 0$. To study this we consider the related optimization problem

$$(17) \quad \begin{array}{ll} \text{minimize} & \langle x, Mx + q \rangle \\ \text{subject to} & Mx + q \geq 0, x \geq 0. \end{array}$$

Given any feasible point x for (17), we define the sets

$$I(x) = \{i \mid M_i x + q_i = 0\} \quad \text{and} \quad J(x) = \{j \mid x_j = 0\}.$$

It is clear that any solution of (17) satisfies

$$I(x) \cup J(x) = \{1, \dots, n\}.$$

We make a convexity(monotone) assumption that M is positive semidefinite and a nondegeneracy assumption that there is a solution of (17), \hat{x} , which satisfies

$$I(\hat{x}) \cap J(\hat{x}) = \emptyset.$$

Under these assumptions, it can be shown that any other solution of (17) satisfies $I(\hat{x}) \subseteq I(x)$ and $J(\hat{x}) \subseteq J(x)$, (see for instance [13, Lemma 2.2]).

THEOREM 3.7. *The solution set of a nondegenerate monotone linear complementarity problem (17) is a set of weak sharp minima for the problem (17).*

Proof. Let x be any solution of (17) and let \hat{x} be the nondegenerate solution. By Theorem 3.2 we must show

$$(\nabla f(\hat{x}))^\perp \cap T(x \mid S) \subseteq \ker(\nabla^2 f(\hat{x})),$$

which, for this problem, means

$$\left\langle (M + M^T)\hat{x} + q, d \right\rangle = 0, \quad \begin{array}{l} M_{I(x)} d \geq 0 \\ d_{J(x)} \geq 0 \end{array} \Bigg\rangle \implies (M + M^T)d = 0.$$

We note that

$$\begin{aligned} 0 &= \langle (M + M^T)\hat{x} + q, d \rangle \\ &= \langle M\hat{x} + q, d \rangle + \langle \hat{x}, Md \rangle \\ &= \sum_{i \in J(\hat{x})} (M\hat{x} + q)_i d_i + \sum_{j \in I(\hat{x})} \hat{x}_j (Md)_j. \end{aligned}$$

Since $I(\hat{x}) \subseteq I(x)$ and $J(\hat{x}) \subseteq J(x)$ and $M_{I(x)} d \geq 0$ and $d_{J(x)} \geq 0$ we see that

$$\sum_{i \in J(\hat{x})} (M\hat{x} + q)_i d_i = 0 \quad \text{and} \quad \sum_{j \in I(\hat{x})} \hat{x}_j (Md)_j = 0.$$

It now follows that $d_{J(\hat{x})} = 0$ and $(Md)_{I(\hat{x})} = 0$ so that $\langle d, Md \rangle = 0$. This is equivalent to $(M + M^T)d = 0$ as required. \square

Note that in this result, we assume that the related optimization problem (17) has a weak sharp minimum, as opposed to an assumption of the form

$$(18) \quad -M\hat{x} - q \in \text{int } N(\hat{x} \mid \mathbf{R}_+^n)$$

as made in [1]. Using Theorem 3.7 it is easy to construct examples that are sharp in the sense given above, but do not satisfy (18).

4. Finite termination of algorithms. In this section we study the convergence properties of algorithms for solving problems of the form

$$(19) \quad \begin{array}{ll} \text{minimize} & f(x) \\ & x \in S \end{array}$$

where it is assumed that $f: \mathbf{R}^n \mapsto \mathbf{R}$ is differentiable and S is a nonempty closed convex subset of \mathbf{R}^n . Under the assumption that the solution set for (19), \bar{S} , is a set of weak sharp minima, we will examine certain tools for identifying an element of \bar{S} in a finite number of iterations. Our approach is based on the techniques developed in [6]. Consequently, we need to introduce some elementary facts concerning the face structure of convex sets.

Recall that a nonempty convex subset \hat{C} of a closed convex set C in \mathbf{R}^n is said to be a face of C if every convex subset of C whose relative interior meets \hat{C} is contained in \hat{C} (e.g., see [21, §18]). In fact, the relative interiors of the faces of C form a partition of C [21, Thm. 18.2]. Thus every point $x \in C$ can be associated with a unique face of C denoted by $F(x|C)$ such that $x \in \text{ri}(F(x|C))$. A face \hat{C} of C is said to be exposed if there is a vector $x^* \in \mathbf{R}^n$ such that $\hat{C} = E(x^* \mid C)$ where

$$E(x^* \mid C) := \arg \max\{\langle x^*, y \rangle : y \in C\}.$$

The vector x^* is said to expose the face $E(x^* \mid C)$. It is well known and elementary to show that every face \hat{C} of a polyhedron is exposed and that the exposing vectors are precisely the elements of $\text{ri}(N(x|C))$ for any $x \in \text{ri } \hat{C}$.

With these notions in mind, we have the following key result.

THEOREM 4.1. *If \bar{S} is a set of weak sharp minima for problem (19) that is regular, then the set*

$$K := \bigcap_{x \in \bar{S}} [T(x \mid S) \cap N(x \mid \bar{S})]^\circ$$

has nonempty interior and for each $z \in \text{int } K$ we have the inclusion $E(z \mid S) \subset \bar{S}$. If it is further assumed that the function f is convex, then \bar{S} is an exposed face of S with exposing vector $-\nabla f(\bar{x})$ for any $\bar{x} \in \bar{S}$.

Proof. The fact that the set K has nonempty interior follows immediately from Corollary 2.7, in particular, $-\nabla f(\bar{x}) \in \text{int } K$ for any $\bar{x} \in \bar{S}$. Let $z \in \text{int } K$ and choose $\delta > 0$ so that $z + \delta B \subset K$. Then for each $\bar{x} \in \bar{S}$

$$\langle z + \delta B, d \rangle \leq 0 \text{ for all } d \in T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}),$$

or, equivalently,

$$\langle z, d \rangle \leq -\delta \|d\| \text{ for all } d \in T(\bar{x} \mid S) \bigcap N(\bar{x} \mid \bar{S}).$$

Hence, given $x \in S$ and $p \in P(x \mid \bar{S})$ we have

$$\langle z, x - p \rangle \leq -\delta \|x - p\|$$

since $(x - p) \in T(p \mid S) \cap N(p \mid \bar{S})$. Consequently, $E(z \mid S) \subset \bar{S}$.

It only remains to show that if f is convex, then $E(-\nabla f(\bar{x}) \mid S) = \bar{S}$ for any $\bar{x} \in \bar{S}$. First observe that

$$(20) \quad \nabla f(x) = \nabla f(y) \text{ for every } x, y \in \bar{S}$$

by Theorem 3.1. Moreover, it has been established that $E(-\nabla f(\bar{x}) \mid S) \subset \bar{S}$. Hence the result will follow if we can show that $\langle \nabla f(x), x \rangle = \langle \nabla f(x), y \rangle$ for any choice of $x, y \in \bar{S}$. But this follows immediately from Theorem 3.1. \square

Remark. In Theorem 4.1, the nonemptiness of the set $\text{int } K$ followed from the differentiability hypothesis on f . In the absence of such a differentiability hypothesis, the result would, in general, be false. Indeed, we need only consider the case $f(x) := \text{dist}(x \mid \bar{S})$, where \bar{S} is any nonempty closed set and S is any set that properly contains \bar{S} in its interior.

In [10], the notion of minimum principle sufficiency was introduced. Assuming $\bar{x} \in \bar{S}$, we define

$$\hat{S} := \arg \min \{ \nabla f(\bar{x})x \mid x \in S \}.$$

Minimum principle sufficiency (MPS) is the equality of the sets \hat{S} and \bar{S} . Note that in the notation above $\hat{S} \equiv E(-\nabla f(\bar{x}) \mid S)$. Solodov [22] pointed out the following interesting corollary to Theorem 4.1.

THEOREM 4.2. *Suppose f is convex and differentiable and $S \subset \mathbb{R}^n$ is a closed convex set. If \bar{S} is a set of weak sharp minima for (19) then MPS is satisfied. If S is polyhedral, then MPS is equivalent to \bar{S} being a set of weak sharp minima for (19).*

Proof. The first statement of the theorem follows from Theorem 4.1. For the second part, note that for all $x \in S$

$$\begin{aligned} f(x) - f(P(x \mid \bar{S})) &\geq \nabla f(P(x \mid \bar{S}))(x - P(x \mid \bar{S})) && \text{by convexity of } f, \\ &= \nabla f(\bar{x})(x - P(x \mid \bar{S})) && \text{by Theorem 3.1,} \\ &= \nabla f(\bar{x})(x - P(x \mid \hat{S})) && \text{by MPS,} \\ &\geq \alpha \|x - P(x \mid \hat{S})\| && \text{by Theorem 3.5,} \\ &= \alpha \|x - P(x \mid \bar{S})\| && \text{by MPS,} \end{aligned}$$

where $\alpha > 0$ as required. \square

The following simple example shows that the assumption of polyhedrality cannot be removed in the above.

Example 4.3. The problem

$$\begin{aligned} &\text{minimize} && x_1 \\ &\text{subject to} && (x_1 - 1)^2 + x_2^2 \leq 1 \end{aligned}$$

has a unique solution $(0, 0)$. It is easy to see that MPS is satisfied. However, for the problem to have a weak sharp minimum would require the existence of $\alpha > 0$ such that

$$x_1 \geq \alpha \sqrt{x_1^2 + x_2^2}$$

for all feasible points x . If we consider points x on the boundary of the circle, then it follows that

$$x_1^2 \geq 2\alpha^2 x_1,$$

which is not true for x_1 sufficiently small.

Another simple application of Theorem 4.1 results in the following strong upper semicontinuity result for linear programs that was first proven in [19, Lemma 3.5].

COROLLARY 4.4. *Let S be a polyhedral convex set in \mathbb{R}^n . Let $c \in \mathbb{R}^n$ and $\bar{S} := \arg \max_{x \in S} \langle \bar{c}, x \rangle$. Then there is a neighborhood U of \bar{c} such that if $c \in U$ then*

$$\arg \max_{x \in S} \langle c, x \rangle = \arg \max_{x \in \bar{S}} \langle c, x \rangle.$$

Proof. If $\bar{S} = \emptyset$, the result follows from the fact that a polyhedral set has a finite number of faces and the graph of the subdifferential of a closed proper convex function is closed. Otherwise, it follows from Theorem 3.5 that \bar{S} is a set of weak sharp minima for

$$\max_{x \in S} \langle \bar{c}, x \rangle.$$

By Theorem 4.1, it follows that $\bar{S} = E(\bar{c} | S)$ and that for all c in a neighborhood of \bar{c} that $E(c | S) \subset E(\bar{c} | S)$. The required equality $E(c | S) = E(c | \bar{S})$ now follows easily. \square

As another immediate consequence of Theorem 4.1, we obtain the following generalization of a result found in [1].

COROLLARY 4.5. *Suppose \bar{S} is a set of weak sharp minima for the problem (19) and let $\{x^k\} \subset \mathbb{R}^n$. If either*

(a) *f is convex and $\{x^k\}$ is any sequence for which $\text{dist}(x^k | \bar{S}) \rightarrow 0$ and ∇f is uniformly continuous on an open set containing $\{x^k\}$, or*

(b) *the sequence $\{x^k\}$ converges to some $\hat{x} \in \bar{S}$, ∇f is continuous and \bar{S} is regular,*

then there is a positive integer k_0 such that any solution of

$$(21) \quad \begin{array}{l} \text{minimize} \quad \langle \nabla f(x^k), x \rangle \\ x \in S \end{array}$$

solves (19).

Proof. Let us first assume that (a) holds. By Theorem 2.6,

$$(22) \quad -\nabla f(\bar{x}) + \alpha B \in \bigcap_{x \in \bar{S}} [T(x | S) \cap N(x | \bar{S})]^\circ$$

for every $\bar{x} \in \bar{S}$, where $\alpha > 0$ is the modulus of weak sharp minimization for the set \bar{S} . Also, by Theorem 3.1, $\nabla f(x) = \nabla f(y)$ for all $x, y \in \bar{S}$. Consequently, the hypotheses imply the existence of an integer k_0 such that $\|\nabla f(x^k) - \nabla f(\bar{x})\| < \alpha$ for all $k \geq k_0$. Therefore, by Theorem 4.1, $E(\nabla f(x^k) | S) = \bar{S}$.

If (b) holds, then (22) is still valid for every point $\bar{x} \in \bar{S}$. The result follows just as it did under assumption (a) since $\|\nabla f(x^k) - \nabla f(\hat{x})\| \rightarrow 0$. \square

The proof of this result only requires the assumption (22) to hold. Part b) of the above corollary can then be proven under the hypothesis that (22) holds only at \hat{x} . This is a weakening of the hypotheses that $-\nabla f(\hat{x}) \in \text{int } N(\hat{x} | S)$ in [1, Thm. 2.1].

Assuming that we can solve (21), Corollary 4.5 can be employed to construct hybrid iterative algorithms for solving problem (19) that will terminate finitely at weak sharp minima. All that needs to be done is to solve the problem (21) occasionally and if an optima is found, then stop. However, some algorithms do not require such a "fix" to locate weak sharp minima finitely. We show that when the objective function f is convex, we can characterize those algorithms that can identify weak sharp minima finitely. We begin with a result that relates the optimality condition given in Theorem 2.6 to the structure of convex subsets of the constraint region S .

LEMMA 4.6. *Let F be any nonempty closed convex subset of the closed convex set $S \subset \mathbb{R}^n$. Then*

$$(23) \quad F + \bigcap_{x \in F} [T(x | S) \cap N(x | F)]^\circ \subset \bigcup_{x \in F} [x + N(x | S)] =: K.$$

Proof. Let $\bar{x} \in F$. We need only show that

$$\bar{K} := \bar{x} + \bigcap_{x \in F} [T(x | S) \cap N(x | F)]^\circ \subset K.$$

Let $y \in \bar{K}$ and let \bar{y} be the projection of $P(y | S)$ onto F . Since $y \in \bar{K}$, there is a $z \in [T(\bar{y} | S) \cap N(\bar{y} | F)]^\circ$ such that $y = \bar{x} + z$. Hence

$$\begin{aligned} 0 &= \langle y - y, P(y | S) - \bar{y} \rangle \\ &= \langle P(y | S) + (y - P(y | S)) - \bar{x} - z, P(y | S) - \bar{y} \rangle \\ &= \langle (P(y | S) - \bar{y}) + (y - P(y | S)) + (\bar{y} - \bar{x}) - z, P(y | S) - \bar{y} \rangle \\ &= \|P(y | S) - \bar{y}\|_2^2 + \langle y - P(y | S), P(y | S) - \bar{y} \rangle \\ &\quad + \langle \bar{y} - \bar{x}, P(y | S) - \bar{y} \rangle + \langle -z, P(y | S) - \bar{y} \rangle. \end{aligned}$$

Observe that each of the terms in the final sum is nonnegative. The second term is nonnegative since $(y - P(y | S)) \in N(P(y | S) | S)$ and $-(P(y | S) - \bar{y}) \in T(P(y | S) | S)$. The third term is nonnegative since $\bar{x} - \bar{y} \in T(\bar{y} | F)$ while $(P(y | S) - \bar{y}) \in N(\bar{y} | F)$. Finally, the fourth term is nonnegative since $(P(y | S) - \bar{y}) \in [T(\bar{y} | S) \cap N(\bar{y} | F)]$. Hence each term is zero so that $\bar{y} = P(y | S)$; that is, $y \in \bar{y} + N(\bar{y} | S) \subset K$. \square

Remarks. 1. It should be noted that one can easily generate examples in which the inclusion (23) is strict.

2. In the fully convex and differentiable case, it was shown in Theorem 4.1 that the set of weak sharp minima \bar{S} is an exposed face of the constraint region S . Consequently, the set F in the above lemma may be taken to be the set \bar{S} . In this case we may write

$$K = \bigcup_{x \in \bar{S}} [F(x | S) + N(x | S)].$$

Lemma 4.6 is now employed to show that the characterization given in [6] of those algorithms that identify the optimal face of S in a finite number of steps also characterizes those algorithms that identify weak sharp minima finitely.

THEOREM 4.7. *Suppose f is convex and let $\bar{S} \subset S$ be a set of weak sharp minima for (19). If $\{x^k\} \subset S$ is such that $\text{dist}(x^k | \bar{S}) \rightarrow 0$ and ∇f is uniformly continuous on an open set containing $\{x^k\}$, then $x^k \in \bar{S}$ for all k sufficiently large if and only if*

$$(24) \quad P(-\nabla f(x^k) | T(x^k | S)) \rightarrow 0.$$

Proof. If $x^k \in \bar{S}$ for all k sufficiently large, then $-\nabla f(x^k) \in N(x^k | S)$ for all k sufficiently large so that (24) holds trivially. On the other hand, suppose (24) is satisfied. The Moreau decomposition of $-\nabla f(x^k)$ yields

$$-\nabla f(x^k) = P(-\nabla f(x^k) | T(x^k | S)) + P(-\nabla f(x^k) | N(x^k | S)).$$

From Theorem 3.1, we have that ∇f is constant on \bar{S} . Thus for any $\bar{x} \in \bar{S}$, the hypotheses imply that

$$\|\nabla f(\bar{x}) + P(-\nabla f(x^k) | N(x^k | S))\| \rightarrow 0,$$

so

$$\text{dist}(x^k + P(-\nabla f(x^k) | N(x^k | S)) | \bar{S} - \nabla f(\bar{x})) \rightarrow 0.$$

However, by Theorem 2.6,

$$\bar{S} - \nabla f(\bar{x}) \subset \text{int} \left[\bar{S} + \bigcap_{x \in \bar{S}} [T(x | S) \cap N(x | \bar{S})]^\circ \right].$$

Thus Lemma 4.6 implies that

$$\begin{aligned} x^k + P(-\nabla f(x^k) | N(x^k | S)) &\in \text{int} \left[\bar{S} + \bigcap_{x \in \bar{S}} [T(x | S) \cap N(x | \bar{S})]^\circ \right] \\ &\subset \bigcup_{x \in \bar{S}} [x + N(x | S)] \end{aligned}$$

for all k sufficiently large. Therefore,

$$\begin{aligned} x^k &= P(x^k + P(-\nabla f(x^k) | N(x^k | S)) | S) \\ &\in P \left(\bigcup_{x \in \bar{S}} [x + N(x | S)] \middle| S \right) \\ &\subset \bigcup_{x \in \bar{S}} \{x\} \\ &= \bar{S} \end{aligned}$$

for all k sufficiently large. \square

In [6], it was shown that the condition (24) is simple to check in certain cases. In particular, it was established that the standard sequential quadratic programming method and the gradient projection method both satisfy (24) and so will automatically generate sequences that terminate finitely at weak sharp minima. We should also note that Polyak[18, Exer. 2, p. 209] indicates that the gradient projection method terminates finitely at weak sharp minima.

REFERENCES

- [1] F. A. AL-KHAYYAL AND J. KYPARISIS, *Finite convergence of algorithms for nonlinear programs and variational inequalities*, J. Optim. Theory Appl., 70 (1991), pp. 319–332.
- [2] L. N. H. BUNT, *Bijdrage tot de Theorie der Convexe Puntverzamelingen*, Thesis, University of Groningen, Groningen, the Netherlands, 1934.
- [3] J. V. BURKE AND M. C. FERRIS, *Characterization of solution sets of convex programs*, Oper. Res. Lett., 10 (1991), pp. 57–60.
- [4] J. V. BURKE, M. C. FERRIS, AND M. QIAN, *On the Clarke subdifferential of the distance function to a closed set*, J. Math. Anal. Appl., 166 (1992), pp. 199–213.
- [5] J. V. BURKE AND S. P. HAN, *A Gauss–Newton approach to solving generalized inequalities*, Math. Oper. Res., 11 (1986), pp. 632–643.
- [6] J. V. BURKE AND J. J. MORÉ, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.
- [7] F. H. CLARKE, *Optimization and Nonsmooth Analysis*. John Wiley, New York, 1983.
- [8] L. CROMME, *Strong uniqueness*, Numer. Math., 29 (1978), pp. 179–193.
- [9] M. C. FERRIS, *Weak sharp minima and penalty functions in mathematical programming*, Tech. Report 779, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, June 1988.
- [10] M. C. FERRIS AND O. L. MANGASARIAN, *Minimum principle sufficiency*, Math. Programming, 57 (1992), pp. 1–14.
- [11] R. HETTICH, *A review of numerical methods for semi-infinite optimization*, in Semi-Infinite Programming and Applications, A. V. Fiacco and K. O. Kortanek, eds., Springer-Verlag, Berlin, 1983.
- [12] K. MADSEN, *Minimization of Non-Linear Approximation Functions*, Dr. Techn. Thesis, Institute for Numerical Analysis, The Technical University of Denmark, Lyngby, Denmark, 1985.
- [13] O. L. MANGASARIAN, *Error bounds for nondegenerate monotone linear complementarity problems*, Math. Programming, 48 (1990), pp. 437–446.
- [14] ———, *A simple characterization of solution sets of convex programs*, Oper. Res. Lett., 7 (1988), pp. 21–26.
- [15] O. L. MANGASARIAN AND R. R. MEYER, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745–752.
- [16] T. S. MOTZKIN, *Sur quelques propriétés caractéristiques des ensembles convexes*, Rend. Accad. Naz. Lincei, 21 (1935), pp. 562–567.
- [17] B. T. POLYAK, *Sharp Minima*, Institute of Control Sciences Lecture Notes, Moscow, USSR, 1979; Presented at the IIASA Workshop on Generalized Lagrangians and Their Applications, IIASA, Laxenburg, Austria, 1979.
- [18] ———, *Introduction to Optimization*, Optimization Software, Inc., Publications Division, New York, 1987.
- [19] S. M. ROBINSON, *Local structure of feasible sets in nonlinear programming, Part II: nondegeneracy*, Math. Programming Stud., 22 (1984), pp. 217–230.
- [20] R. T. ROCKAFELLAR, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.
- [21] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [22] M. V. SOLODOV, Private communication, 1992.

AN INTERIOR-POINT METHOD FOR MINIMIZING THE MAXIMUM EIGENVALUE OF A LINEAR COMBINATION OF MATRICES*

FLORIAN JARRE†

Abstract. An algorithm for minimizing the largest eigenvalue of an affine combination of symmetric matrices is presented. The nonsmooth problem is transformed into an equivalent smooth constrained problem, which is solved by a predictor-corrector interior-point method taking full advantage of the differentiability and convexity. Some promising numerical results obtained from a preliminary implementation are included.

Key words. convex program, eigenvalue, implementation

AMS subject classifications. 65FF99, 65K10, 90C25

1. Introduction.

1.1. The problem. Problems in mechanics and systems analysis can often be expressed in the following (or a closely related) form: Find

$$(1) \quad \lambda^{\text{opt}} := \inf_{x \in \mathbb{R}^m} \{ \lambda_{\max}(A(x)) \}$$

and an optimal solution x^{opt} if it exists. Here $\lambda_{\max}(A(x))$ is the largest eigenvalue of the matrix

$$(2) \quad A(x) := A^{(0)} + \sum_{i=1}^m x_i A^{(i)},$$

for given symmetric matrices $A^{(i)} \in \mathbb{R}^{n \times n}$, $0 \leq i \leq m$.

Since $\lambda_{\max}(A(x))$ is convex, problem (1) is a convex but nondifferentiable optimization problem. It arises in a variety of applications; for example, the stability analysis of dynamical systems [2], [3], or in combinatorial applications [1]; see also [5], [26], [28] for further applications.

Below, we state some simple facts that illustrate the nature of problem (1) and are easily verified.

- If $\lambda^{\text{opt}} > -\infty$, it lies in the interval $[\lambda_{\min}(A^{(0)}), \lambda_{\max}(A^{(0)})]$,
- There is no x such that $\sum_i x_i A^{(i)}$ is positive definite if and only if λ^{opt} is finite,
- There is no $x \neq 0$ such that $\sum_i x_i A^{(i)}$ is semidefinite if and only if problem (1) has a nonempty and bounded set of optimal solutions

1.2. Basic assumptions. Without loss of generality, we assume that the matrices $I, A^{(i)}$, $0 \leq i \leq m$ are linearly independent, i.e., $r_0 I + \sum_{i \geq 1} r_i A^{(i)} = 0$ if and only if $r_i = 0$, $0 \leq i \leq m$. We further assume that $\lambda_{\min}(A^{(0)}) = 0$ (by adding a suitable multiple of the identity to $A^{(0)}$). For convenience, we also assume that all $A^{(i)}$ have Frobenius norm 1.

* Received by editors July 29, 1991; accepted for publication (in revised form) June 2, 1992.

† Institut für Angewandte Mathematik, University of Würzburg, 8700 Würzburg, Germany (jarre@vax.rz.uni-wuerzburg.dbp.de). This work was done at the Department of Operations Research at Stanford University, Stanford, California, 94305.

2. Relationship to other methods. The need for efficient methods for solving minimax eigenvalue problems like (1) is expressed, for example, in [28]. In this section, we briefly introduce different methods for solving (1) and outline the development that led to the method of this paper.

2.1. “Classical” methods. Standard methods for nondifferentiable optimization, e.g., Shor’s subgradient method, Shor’s ellipsoid method, or Kelly’s cutting plane method, can be used for solving (1); for a survey of these methods, refer to [2, Chaps. 13–14]. Other papers on methods for nondifferentiable optimization include [17], [21], [31], [38]. Specialized methods for (1) are developed in, e.g., [3], [8], [14], [15], [25], [26], [27], [29].

Overton’s method [25], [26] is locally quadratically convergent. His method builds on earlier work of Fletcher [5] for semidefinite programming. Goh and Teo’s method [8] is most closely related to the approach of this paper, since it involves transforming (1) into a smooth nonlinear programming problem by means of determinants.

2.2. Development of interior-point methods. The present paper was motivated by the recent developments of interior-point methods that started with the work of Karmarkar [16] in 1984. Karmarkar presented a method for solving linear programs. Both his proof showing that the method was polynomial (and of much lower complexity than the ellipsoid method) as well as his claim that an implementation of his method was superior to the programs based on the simplex method received great attention (and skepticism). It was soon recognized that Karmarkar’s original method could easily be modified (and improved) to handle nonlinear convex constraints (see Sonnevend [32]), and analyses that proved that the rate of convergence for programs with certain nonlinear convex constraints is the same as in the case of linear constraints were given in [9], [10], [19].

Of particular interest for this paper is the application of interior-point methods for optimization problems over the cone of positive semidefinite matrices, as suggested by Sonnevend [32], [33] (implementational aspects were not considered). In [34] he compared different barrier functions for the cone of positive definite matrices and showed why the barrier function $-\log \det X$ is the “principal actor” for such optimization problems. A detailed complexity analysis of interior-point methods for convex optimization problems based on the notion of self-concordance is given in Nesterov and Nemirovsky’s work [22]. Their analysis also applies to the function $-\log \det X$ and provides the main theoretical tools for our results in §3.

During the writing of this paper, we learned of two other papers [1], [30] that also present implementations of interior-point methods for solving eigenvalue problems and are closely related to this one. In [1] Alizadeh gives an elegant proof for a duality result concerning minimax eigenvalue problems and applies this result to solve combinatorial problems. His method is based on the same theoretical framework as ours, and solves the dual problem. It is particularly attractive for the case where $m \gg n$. In [30] Ringertz describes an implementation of a barrier method for minimizing the largest eigenvalue of a matrix subject to nonlinear equality and inequality constraints. His problems arise in the optimal design of shell structures for airplanes. They are not convex and hence more difficult in nature than problem (1). Ringertz does not touch the topic of estimating the complexity of his method but presents numerical experiments on a Cray XMP with several thousand nonlinear constraints. His results support the obvious conjecture (see also [12]) that interior-point methods are also very efficient for much more difficult problems than the one analyzed here and for which a complexity analysis is not yet possible. Nevertheless, the complexity analysis

of interior-point algorithms is an integral part of the method, not only since it led to their development, but also since it explains their behaviour (weak dependence on the data, affine invariance) and gives important conclusions for implementational aspects (line search, stopping test).

As in [8], problem (1) is transformed to a nonlinear optimization problem that is solved by an interior-point method taking full advantage of the structure (differentiability and convexity) of the transformed problem. The efficiency of the interior-point method is what makes the approach attractive. The algorithm presented here is globally linearly convergent and the guaranteed rate of convergence depends only on the dimension of the problem and not on the data of the matrices $A^{(i)}$. More precisely, there is a theoretical worst-case bound of at most $O(\sqrt{n})$ iterations to reduce the error $\lambda_{\max}(A(x)) - \lambda^{\text{opt}}$ by a factor of 2. Furthermore, recent analyses of interior point methods for linear programs by Ye et al. in [37] showed that predictor-corrector methods similar to the one presented here are quadratically convergent under weak assumptions.

2.3. Implementations of interior-point methods. An early implementation of a barrier method, the SUMT program [4] by Fiacco and McCormick, did not prove to be competitive, and barrier methods were subsequently largely ignored. The recent theoretical developments led to a deeper understanding of the method, and soon after Karmarkar's original work, a number of implementations of his method and its variants for solving linear programs were presented. For certain problems, these implementations were clearly superior to the simplex method. However, numerical implementation of interior-point methods for solving nonlinearly constrained problems has been very slow. A simple program for problems with quadratic constraints is presented in [13]. Recently, Nesterov and Nemirovsky also distributed a software package [24] with an implementation of an interior-point method for solving a class of convex programming problems, including problem (1). Their package is based on a projective algorithm that is closely related to Karmarkar's original algorithm. The package is easy to use, but it is quite sensitive to ill conditioning, even for the small test problems provided with the package (see §5.3). As motivated in §4, the implementation of this paper may be attractive for the case where n is large and m is moderate, and the matrices A_i are sparse or of low rank. Very promising also is the recent implementation of Alizadeh [1] (for large m and moderate n) and the one by Ringertz [30] that is tailored to the solution of more general problems.

3. General barrier approach. It is our goal to present a fairly self-contained article that provides some encouragement (supported by numerical experiments) that interior-point methods are an ideal tool for solving minimax eigenvalue problems. However, the algorithm described below requires improvement (to exploit sparsity or low-rank structure of the matrices $A^{(i)}$ and to allow for more general problems) to be really efficient.

3.1. An equivalent problem. We rewrite problem (1) as another convex differentiable problem with a positive definite constraint. In the following, we write $A > 0$ ($A \geq 0$) if the matrix A is positive definite (positive semidefinite). Positive definiteness can easily be verified numerically (via Cholesky decomposition). Moreover, the cone of positive definite matrices $A > 0$ is characterized by the domain of the smooth convex barrier function $-\log \det_+(A)$. Here, \det_+ of a symmetric matrix

is simply defined as

$$(3) \quad \det_+(A) := \begin{cases} \det(A) & \text{if } A \text{ is positive definite,} \\ -\infty & \text{otherwise.} \end{cases}$$

Let us introduce an $(m+1)$ th variable λ and denote

$$(4) \quad \tilde{x} := (x^T, \lambda)^T \in \mathbb{R}^{m+1}.$$

Note that a number λ is larger than $\lambda_{\max}(A(x))$ if and only if $\lambda I - A(x)$ is positive definite. This observation can be used to rewrite (1) and (2) as

$$(5) \quad \inf_{\tilde{x} \in \mathbb{R}^{m+1}} \{\lambda \mid \det_+(\lambda I - A(x)) > 0\}.$$

The domain of problem (5) is denoted by the open set S° . The efficiency of a barrier method for solving (5) as given below strongly depends on the properties of the barrier function for S° . We use the barrier function

$$(6) \quad \phi(\tilde{x}) := -\log \det_+(\lambda I - A(x)).$$

Under the basic assumptions of §1, we can further show that this function is strictly convex.

LEMMA 1. *The function $-\log \det_+(\lambda I - A(x))$ is strictly convex in x, λ .*

Proof. For the proof, see Appendix A. \square

3.2. A conceptual barrier method. In the following, we briefly recall the principle of a barrier method. Modifications like bounds on the variables, linear equality constraints, or additional convex constraints are easy to implement but are omitted here for clarity.

Suppose that a function $f_0(\tilde{x})$ (which in our case is simply the $(m+1)$ th variable λ) is to be minimized over some convex feasible set S (e.g., the set where $\lambda I - A(x)$ is positive semidefinite), and we are given a “barrier function” for the set S (i.e., a smooth convex function $\phi(\tilde{x})$ that tends to infinity as x approaches the boundary of S and is finite in its interior S°).

We consider a family of unconstrained subproblems. For $\mu \rightarrow 0$ ($\mu > 0$), find

$$(7) \quad \tilde{x}(\mu) := \arg \min \{f_0(\tilde{x}) + \mu \phi(\tilde{x})\}.$$

The minimizers $\tilde{x}(\mu)$ are unique if, for example, S is bounded, ϕ is strictly convex, and f_0 is convex. As the perturbation $\mu \phi(\tilde{x})$ of the objective function is “phased out,” i.e., as $\mu \rightarrow 0$, we can show under weak assumptions that $\tilde{x}(\mu)$ converges to an optimal solution of the original problem. Moreover, $\tilde{x}(\mu)$ is a smooth curve and the tangent to $\tilde{x}(\mu)$ is easily computable; see, e.g., [4]. These properties motivate the following conceptual barrier method.

Given $\mu_0 = 1$ and $\tilde{x}(\mu_0)$, set $k := 0$

Do until convergence

- Compute the tangent $\tilde{x}'(\mu_k)$.
- Select an appropriate $\mu_{k+1} < \mu_k$.
- Predict the next iterate by $\hat{x}(\mu_{k+1}) := \tilde{x}(\mu_k) + (\mu_{k+1} - \mu_k)\tilde{x}'(\mu_k)$.
- Find $\tilde{x}(\mu_{k+1})$ by Newton’s method starting from $\hat{x}(\mu_{k+1})$.
- Set $k := k + 1$.

End

3.3. Barrier methods and self-concordance. The reason we expect barrier methods for problem (5) to work well depends on the properties of the barrier function ϕ . In particular, the property of *self-concordance*, which is defined by Nesterov and Nemirovsky in [23], plays an important role for the analysis of Newton's method. In its simplest form, the condition that a function ϕ be self-concordant is as follows.

DEFINITION. Let $S \subset \mathbb{R}^n$ be convex, $x \in S^\circ$, and $h \in \mathbb{R}^n$ be arbitrary and define the restriction of a convex function $\phi : S^\circ \rightarrow \mathbb{R}$ in direction h through x by the function $\psi : J \rightarrow \mathbb{R}$, where $\psi(t) := \psi_{x,h}(t) := \phi(x + th)$ and J is some open interval containing 0. The function ϕ is called self-concordant, with self-concordance parameter α if, for any x and h , its restriction ψ satisfies the inequality

$$(8) \quad |\psi'''(0)| \leq \frac{2}{\sqrt{\alpha}} \psi''(0)^{3/2}.$$

A self-concordant function $\phi : S^\circ \rightarrow \mathbb{R}$ is called *strongly self-concordant* if $\phi(x) \rightarrow \infty$ as x approaches the boundary ∂S of S . Thus self-concordant barrier functions are strongly self-concordant.

Discussion. Intuitively, this condition may be interpreted as follows. The third derivative ψ''' is a measure of how fast the second derivative ψ'' is changing, and this condition implies that the *relative* change of ψ'' is locally¹ bounded by some Lipschitz constant $2/\sqrt{\alpha}$. Since the condition holds for any $h \in \mathbb{R}^n$, an analogous statement also holds for the second and third derivatives $D^2\phi$ and $D^3\phi$ of ϕ at x .

For the derivation of an equivalent condition, a “relative Lipschitz condition” on the Hessian of ϕ , and for a simplified analysis we refer to [10], [11].

Nesterov and Nemirovsky showed in [23] that this condition is satisfied by a class of *logarithmic* barrier functions. In particular, it is satisfied with $\alpha = 1$ for the barriers of linear or convex quadratic constraint functions and the function “ $-\log \det_+(A)$.” They further show that, if some function Θ is self-concordant on a convex subset of \mathbb{R}^l and if L is an affine mapping into \mathbb{R}^l (its range intersecting at least part of the domain of Θ), then $\phi(y) := \Theta(L(y))$ is also self-concordant. This implies the following lemma.

LEMMA 2. *The barrier function $-\log \det_+(\lambda I - A(x))$ is self-concordant in x, λ with $\alpha = 1$.*

Results. Denote the Hessian of ϕ by $H(x) := D^2\phi(x)$. The main result about self-concordance in [23] is that Newton's method when applied to minimizing a strongly self-concordant function ϕ defined on a set S° is quadratically convergent (with constant less than 2) if the first Newton step Δx exists and has length less than $\sqrt{\alpha}/4$ measured in the “right” norm

$$(9) \quad \|\Delta x\|_{H(x)} := (\Delta x^T H(x) \Delta x)^{1/2} = \psi''_{x,\Delta x}(0)^{1/2}.$$

The importance of this statement is that it only depends on α and that the set of points x for which the Newton step Δx (starting at x) satisfies $\|\Delta x\|_{H(x)} \leq \sqrt{\alpha}/4$ is a “fixed portion” of the whole set S . More precisely, let ϕ be a self-concordant logarithmic barrier function of some convex constraint functions² f_1, \dots, f_m ,

$$\phi(x) = - \sum_{i=1}^m \log(-f_i(x));$$

¹ Here, “locally” means that $t\psi''(0)^{1/2}$ is small.

² Note that $-\sqrt[n]{\det_+(\lambda I - A(x))}$ is convex; so this statement applies to the function $\phi(\bar{x}) = -\frac{1}{n} \log \det_+(\lambda I - A(x))$ with $m = 1$ and $\alpha = 1/n$.

let the domain $S = \{x \mid f_i(x) \leq 0\}$ of ϕ be bounded, \bar{x} be the minimum of ϕ , and $E = 0.2\{h \mid h^T D^2 \phi(\bar{x}) h \leq 1\}$. Then, $\bar{x} + E$ is within the domain of quadratic convergence of Newton's method and

$$(10) \quad \bar{x} + 5\sqrt{\alpha}E \subset S \subset \bar{x} + \frac{5(m+2\sqrt{m})}{\sqrt{\alpha}}E$$

(see [11] and also [35] for a similar result for quadratic constraints). Applied to the set S of problem (1), this implies that the ratio of inner and outer ellipsoid only depends on the dimension of the matrices $A^{(i)}$ but not on the matrices themselves. This statement implies further that all subproblems (7) from §3.2 are of the same “difficulty”; that is, no matter how small μ is, the domain of convergence of Newton's method for finding $\tilde{x}(\mu_{k+1})$ is always a “fixed percentage” of the previous level set $\{\tilde{x} \mid \tilde{x} \in S, f_0(\tilde{x}) \leq f_0(\tilde{x}(\mu_k))\}$, and the “percentage” depends only on n and μ_k/μ_{k+1} . This fact guarantees polynomial convergence (independent of the data matrices $A^{(i)}$) of the barrier method in §3.2, and it does not hold for other barrier functions such as $1/\det(A(x))$.

4. A specific barrier method.

4.1. Notation and initialization. The ideas outlined above will now be applied to problem (5) in more detail. Recall the definition (4) of \tilde{x} and define

$$(11) \quad A(\tilde{x}) := \lambda I - A(x) = \lambda I - A^{(0)} - \sum_{i=1}^m x_i A^{(i)}.$$

For a given vector \tilde{w} in \mathbb{R}^{m+1} and for $\mu \in (0, 1]$, consider the barrier function

$$(12) \quad \varphi(\tilde{x}, \mu) := \frac{\rho}{\mu} \lambda - \log \det_+(A(\tilde{x})) - \tilde{w}^T \tilde{x}$$

with some constant $\rho > 0$. Note that the feasible domain of (5) is given by the domain of φ , $S^\circ = \{\tilde{x} \mid \varphi(\tilde{x}, \mu) < \infty\}$.

The vector \tilde{w} is a linear perturbation to the barrier function and does not influence self-concordance of φ . We will choose \tilde{w} so that our initial point is a minimizer of $\varphi(\cdot, 1)$.

To initialize the algorithm, let $\lambda_0 = 2\lambda_{\max}(A^{(0)}) > 0$ (recall $\lambda_{\min}(A^{(0)}) = 0$) and let $x^0 = 0 \in \mathbb{R}^m$. This implies that $\tilde{x}^0 = (0, \dots, 0, 2\lambda_{\max}(A^{(0)}))^T$ is a strictly feasible starting point (and that $\text{cond}(A(\tilde{x}^0)) = 2$). Set $\mu = 1$ and choose ρ as $\rho = n/\lambda_{\max}(A^{(0)})$ for example. The vector \tilde{w} is chosen such that

$$(13) \quad D_{\tilde{x}} \varphi(\tilde{x}^0, 1) = 0$$

and is constant throughout our algorithm. Let $e_{m+1} := (0, \dots, 0, 1)^T \in \mathbb{R}^{m+1}$. Then another representation of φ is given by

$$(14) \quad \varphi(\tilde{x}, \mu) = \left(\frac{\rho}{\mu} e_{m+1} - \tilde{w} \right)^T \tilde{x} - \log \det_+(A(\tilde{x})).$$

4.2. The derivatives of φ . Adopting the notation in [1], we define the “symmetric scalar product” of two matrices by

$$(15) \quad A \bullet B := \sum_{i,j} A_{ij} B_{ij} = \text{trace}(A^T B).$$

The first and second derivatives of φ are defined in terms of

$$(16) \quad B(\tilde{x}) := A(\tilde{x})^{-1}.$$

To abbreviate the notation, we refer to $\tilde{x}_{m+1} = \lambda$ and $A^{(m+1)} = -I$. It is known (and easy to verify) that the partial derivatives of “log det” are

$$\frac{\partial}{\partial X_{ij}} \log \det(X) = (X^{-T})_{ij}$$

(for $X \in \mathbb{R}^{n \times n}$ with $\det(X) > 0$ [6]), so that the derivative $D \log \det X = X^{-T}$, where X^{-T} is to be interpreted as a linear functional via the scalar product (15), $X^{-T}[A] := X^{-T} \bullet A$. Thus, by the chain rule, we obtain

$$\hat{g}_i := \hat{g}_i(\tilde{x}) := \frac{\partial}{\partial \tilde{x}_i} \log \det_+(A(\tilde{x})) = B(\tilde{x}) \bullet \left(\frac{d}{d\tilde{x}_i} A(\tilde{x}) \right).$$

Hence

$$(17) \quad \hat{g}_i = -B(\tilde{x}) \bullet A^{(i)}, \quad 1 \leq i \leq m+1.$$

The derivative $g(\tilde{x}, \mu) = D_{\tilde{x}} \varphi(\tilde{x}, \mu)^T$ of φ now follows from (14):

$$(18) \quad g(\tilde{x}, \mu) = \frac{\rho}{\mu} e_{m+1} - \hat{g}(\tilde{x}) - \tilde{w}.$$

The derivative $D_{\tilde{x}} g(\tilde{x}, \mu)$ of g , in turn, is given by the partial derivatives

$$\frac{\partial}{\partial \tilde{x}_j} (A^{(i)} \bullet B(\tilde{x})) = \sum_{k,l} A_{kl}^{(i)} \frac{\partial}{\partial \tilde{x}_j} B_{kl}(\tilde{x}) = \sum_{k,l} A_{kl}^{(i)} (-B_{kq} B_{rl})_{qr} \bullet \left(\frac{\partial}{\partial \tilde{x}_j} A(\tilde{x}) \right)$$

for $1 \leq i, j \leq m+1$. (Here and in the following, we occasionally omit the argument of $B = B(\tilde{x})$.) Summarizing, we find that

$$\frac{\partial}{\partial \tilde{x}_j} g_i(\tilde{x}, \mu) = \sum_{k,l,q,r} A_{kl}^{(i)} A_{qr}^{(j)} B_{kq} B_{rl} = B A^{(i)} \bullet A^{(j)} B = \text{trace}(A^{(i)} B A^{(j)} B).$$

It is straightforward to compute the Hessian

$$(19) \quad H(\tilde{x}) := D_{\tilde{x}} g(\tilde{x}, \mu) = \left(\frac{\partial}{\partial \tilde{x}_j} g_i(\tilde{x}, \mu) \right)_{ij}$$

in $O(mn^3 + m^2n^2)$ multiplications. For our examples below, with $m, n < 100$, this was sufficient; we note, however, that, if the $A^{(i)}$ are sparse or of low rank (and n is large, while m is moderate), this can be done much more efficiently. In this case, the Hessian is a dense (m by m) matrix, even if the (n by n) matrix $A(\tilde{x})$ is sparse. Also, note that the Hessian does not depend on μ , ρ , or \tilde{w} .

4.3. Perturbed centers.

DEFINITION. A minimizing point of the function $\varphi(\cdot, \mu)$ will be called the *perturbed center* is denoted by $\tilde{x}(\mu)$. If it exists, the perturbed center with perturbation $\tilde{w} = 0$ is the *analytic center*, as defined in [32].

LEMMA 3. *The following conditions hold:*

(i) Let some point $x^0 \in S^o$ be given and \tilde{w} be such that $D\varphi(x^0, 1) = 0$. Then the functions $\varphi(\cdot, \mu)$ have unique minimizers $\tilde{x}(\mu)$ for all $\mu \in (0, 1]$ if and only if λ^{opt} is finite;

(ii) The analytic center exists for $\mu = 1$ if and only if it exists for all $\mu \geq 0$, or, equivalently, if and only if the set of optimal solutions for (1) is nonempty and bounded.

Proof. For the proof, see Appendix A. \square

(Note that, if λ^{opt} is finite, the set of optimal solutions for (1) can be empty or unbounded.)

It is intuitive (and has been shown, e.g., in [4], [12]) that, if (1) has an optimal solution, the perturbed centers $\tilde{x}(\mu)$ converge to a solution \tilde{x}^{opt} as $\mu \rightarrow 0$. This result holds for any strictly feasible starting point \tilde{x}^0 and the corresponding \tilde{w} . Thus Lemma 3 describes a whole family of curves of perturbed centers, each of which starts somewhere in S^o and leads to an optimal solution.

For the analytic center, a number of nice properties can be shown. In particular, the center is affine invariant and “far away” from the boundary of the feasible set in the sense that for $\tilde{w} = 0$ the points $\tilde{x}(\mu)$ allow two-sided ellipsoidal approximations of the level sets $\{\tilde{x} \in S \mid f_0(\tilde{x}) \leq \tau\}$ for a suitable $\tau = \tau(\mu)$ as already stated in (10). The inner ellipsoid also describes the domain in which the path of centers $\tilde{x}(\mu)$ is well approximated by its tangent. Furthermore, points on the path of analytic centers define dual feasible variables that can be used in a stopping test. Similar results hold in a somewhat weaker form for the perturbed center if the perturbation \tilde{w} is small. For large perturbations \tilde{w} (for $\tilde{w}^T D_{\tilde{x}}^2 \varphi(\tilde{x}^0, 1)^{-1} \tilde{w} \geq 1$), however, this is no longer true. The vector \tilde{w} guarantees the existence of the perturbed center and that the method below is well defined even if the set of optimal solutions is unbounded. However, it is our goal to keep \tilde{w} small. To decrease the norm of \tilde{w} , it is advantageous in some instances to perform a phase one and to introduce an additional bound, for example, $\|\tilde{x}\|_2 \leq 10^6$. Then find the *analytic* center of this set (the existence of it is guaranteed by the additional bound) and define \tilde{w} as in (13) for this starting point after deleting the additional bound again.

The algorithm in this paper is a modification of that described in [12]. The situation here is particularly simple, since a strictly feasible starting point can easily be constructed. Below, we explain one iteration of the method.

4.4. One iteration of the algorithm. The following description is a more detailed and specialized version of the barrier method outlined in §3. At the beginning of each iteration, we assume that a strictly feasible current iterate \tilde{x}^k and an associated parameter μ_k are given, where \tilde{x}^k is considered as an approximation to $\tilde{x}(\mu_k)$. We repeat the following iteration until some convergence criterion is satisfied.

1) Determine the tangent to the curve of perturbed centers $\tilde{x}^k(\mu)$, that passes through \tilde{x}^k . The tangent to the curve $\tilde{x}(\mu)$ is easily obtained by differentiating the characteristic equation

$$(20) \quad g(\tilde{x}(\mu), \mu) \equiv 0$$

with respect to μ . The differentiation yields

$$D_{\tilde{x}} g(\tilde{x}(\mu), \mu)^T \tilde{x}'(\mu) + \frac{\partial}{\partial \mu} g(\tilde{x}(\mu), \mu)^T = H(\tilde{x}(\mu)) \tilde{x}'(\mu) - \frac{\rho}{\mu^2} e_{m+1} = 0.$$

The tangent can thus be computed from a Cholesky decomposition of the Hessian. Note that the direction of the tangent $\tilde{x}'(\mu)$ does not depend on the vector \tilde{w} or on

μ , but only on the current point \tilde{x} . Thus the above computation precisely yields the direction of the tangent to the (unique) curve of perturbed centers starting at \tilde{x} and ending in an optimal solution. Also, the last component (i.e., λ) of $\tilde{x}'(\mu)$ is positive (since H is positive definite), so that $-\tilde{x}'(\mu)$ is obviously a descent direction for the objective value λ .

2) Perform a linear extrapolation in the direction of the tangent, as follows:

$$\hat{x}^{k+1} := \tilde{x}^k - \beta_k \tilde{x}'(\mu).$$

The steplength β_k for the extrapolation is chosen as r_k times the maximum steplength β_{\max} such that $A(\tilde{x}^k - \beta_{\max} \tilde{x}'(\mu))$ is still positive semidefinite. (The exact computation of β_{\max} involves computing the maximum eigenvalue of a matrix.) Here, $r_k \in (0, 1)$ can be chosen adaptively: Set $r_1 = 0.9$, and, (for $k \geq 2$) if Newton's method (in step 4), below) converged in less than three steps in the previous iteration, increase r (as $r_k = (1 + r_{k-1})/2$); if it converged in more than four Newton steps, decrease r (as $r_k = \max\{r_{k-1}/2, 2r_{k-1} - 1\}$).

3) The parameter μ is controlled as follows. Given the k th iterate \tilde{x}^k , an associated parameter μ_k , and an extrapolated point $\hat{x}^{k+1} \in \mathbb{R}^{m+1}$ for the $(k+1)$ th iterate, determine the (unique) value $\hat{\mu}_{k+1}$ that “best fits” \hat{x}^{k+1} , i.e., such that the last component of the gradient g at \hat{x}^{k+1} is zero. Set $\mu_{k+1} = \min\{\mu_k/2, \hat{\mu}_{k+1}\}$.

Remark. For a moderate stepsize β for the (linear) extrapolation, we can show that $\hat{\mu}_{k+1} < \mu_k$ (for a more precise statement and a proof, we refer to Appendix A).

4) Perform Newton's method (with line search) to minimize the barrier function $\varphi(\cdot, \mu_{k+1})$. The stopping criterion for Newton's method is $\|\Delta\tilde{x}\|_{H(\tilde{x})} \leq 0.2$, i.e., the H -norm of the Newton step (9) must be less than 0.2.

In our program, below, we implement a simple modification of Newton's method by using the “old” factorization of the Hessian that was computed for the last Newton step (or in the extrapolation) for four further “inexact” Newton steps with line search before recomputing the new Hessian for the next true Newton step.

4.5. Stopping test. The following lemma may be used for determining when to stop.

LEMMA 4. *As in (9), we define, for a positive definite matrix H , the H -norm of a vector v by $\|v\|_H^2 = v^T H v$. If, for a given point \tilde{x}^k , there is a μ such that the Newton step $\Delta\tilde{x}$ for finding the analytic center $\tilde{x}(\mu)$ (with $\tilde{w} = 0$) satisfies*

$$\|\Delta\tilde{x}\|_H \leq \frac{\sqrt{\alpha}}{4}$$

with $H := H(\tilde{x}^k)$, then (1) has a bounded optimal solution set, and \tilde{x}^k is almost optimal in the following sense: $\lambda - \lambda^{\text{opt}} \leq 2n\mu/\rho$.

The assumption that there is such a μ can be interpreted as \tilde{x}^k being moderately close to the path of analytic centers.

4.6. Ill-conditioning. The barrier method of the previous sections encounters ill-conditioned matrices (as μ tends to zero). We briefly discuss the two critical points in our computations.

1) First, as μ tends to zero, the matrix $A(\tilde{x})$ approaches the singular matrix $A(\tilde{x}^{\text{opt}})$, and is therefore ill conditioned. Simple considerations show that the minimum eigenvalue of $A(\tilde{x}(\mu))$ is of order $O(\mu)$. It is our experience that for the small examples below with $\mu \geq 10^{-10}$ the inversion of the matrix $A(\tilde{x})$ can be carried out to sufficient accuracy (Matlab is using double precision arithmetic), provided that the

maximum eigenvalue of $A(\tilde{x}^{\text{opt}})$ is $O(1)$. For many practical applications, it suffices to approximate the optimal solution up to low accuracy only (e.g., $\mu \geq 10^{-6}$), and, in these cases, we expect the barrier method to work well even for larger problems. We do not examine any methods for stabilizing the computation of $A(\tilde{x})^{-1}$ in the case that a very high accuracy of the optimal solution is desired, and we suspect that further difficulties may arise.

However, interior-point methods that use a centering procedure at each iteration have an “implicit stability” that can be explained as follows: From the proof of Lemma 1 (Appendix A), it is obvious that the evaluation of the Hessian yields a positive (semi)definite matrix H , regardless of how inaccurate the computation of the matrix $B = A(\tilde{x})^{-1}$ is. In our algorithm, we use a uniformly positive definite approximation \hat{H} of the Hessian H that is obtained from a simple “stabilization procedure,” described below. Since inexact Newton’s method with a line-search converges for *any* uniformly positive definite approximation \hat{H} of the Hessian, the center will be found eventually, even if H is a random positive definite matrix. Similarly, the extrapolation is a descent direction for the objective value λ for *any* positive definite approximation of H (see step 1) of §4.4). This property suffices to prove that a method as outlined above always converges to an optimal solution \tilde{x}^{opt} , even if the computation of the Hessian is inaccurate. We note that such a statement is not true for methods that do not use a centering procedure. If the Hessian is computed inaccurately, such a method may converge to a nonoptimal point. Furthermore, in general, the Hessians that are evaluated at points on the central path are less ill conditioned than the Hessians that are evaluated at points far away from the central path. (For barrier functions of linear programs with a unique optimal solution, we can prove, for example, that the condition number of the Hessians near the central path is globally bounded independent of μ , while, in general, the condition number of the Hessian becomes unbounded as the argument approaches the boundary of the feasible set.) These facts may also explain our observation in §5.3.

2) Second, as μ tends to zero, the Hessian $H(\tilde{x}(\mu)) = D^2\varphi(\tilde{x}(\mu), \mu)$ becomes ill conditioned. This type of ill conditioning also occurs when solving (degenerate) linear programs by interior-point methods. Lustig, Marsten, and Shanno [18], Gill et al. [7], and others present solutions for such instabilities (via iterative refinement) and show with a large number of numerical experiments that their strategies are effective. In our program, we use the following simple strategy: Given the Hessian $H = H(\tilde{x})$ at some point \tilde{x} , set $\hat{H} := H + \epsilon \|H\|_{\infty} I$ (with $\epsilon = 10^{-10}$) and denote the Newton step by $\Delta\tilde{x} = H^{-1}g$ and the “stabilized Newton step” by $\Delta\hat{x} = \hat{H}^{-1}g$. As motivated in definition (9), the canonical norm associated with this method at the point \tilde{x} is the norm $\|h\|_H = (h^T H h)^{1/2}$. It is straightforward to verify (e.g., Stewart and Sun, [36, Thm. 2.11, p. 124]) that

$$\frac{\|\Delta\hat{x} - \Delta\tilde{x}\|_H}{\|\Delta\tilde{x}\|_H} \leq \epsilon \frac{\text{lub}_{\infty}(H)}{\text{lub}_2(H^{-1})} \leq \sqrt{n} \epsilon \text{cond}_2(H).$$

Since $\Delta\hat{x}$ is just an iterative direction for a line-search during Newton’s method, this error is not crucial. For very ill-conditioned H , though, this stabilization does increase the number of Newton steps.

5. Some numerical results. The method outlined in §4 has been implemented in Matlab [20]. We present a few preliminary examples for which the method shows the same convergence behavior as interior-point methods applied to linear programs. This—along with the work of Alizadeh (see Appendix B of this paper)—strongly sug-

TABLE 1

n	m	$s_{m,n}$	Iterations	Hessians	Inexact Newton steps	Multiplicity
5	10	500	7.8/9	9.5/14	23.7/37	3.0
10	20	100	9.0/11	11.4/15	29.6/42	3.8
20	40	100	10.0/12	12.5/15	33.2/43	5.1
40	80	20	11.1/13	14.0/16	34.4/42	6.7

gests that the method may be very efficient for solving minimax eigenvalue problems and for more general convex programming problems (see [12]).

5.1. Random examples. In Table 1, we give results for minimizing the largest eigenvalue of affine combinations of m random symmetric matrices of size n for different values of m and n . Our stopping criterion is $n\mu/\rho \leq 10^{-4}$. In each iteration, we performed one extrapolation and a small number of Newton and “inexact Newton” steps with a line search. We list m , n , and the sample size $s_{m,n}$ of how many random problems were tested for each pair m, n . We further list the average and the maximum number of iterations needed and the average and maximum number of evaluations of the Hessian and total (inexact) Newton steps. Finally, we record the average multiplicity of the maximum eigenvalue in the optimum. For $n = 5$, only about 20% of all problems had a finite λ^{opt} . For the unbounded problems, a negative definite combination was found in the first iteration³, so that we only list the statistics for the bounded problems. For $n = 10$, four problems were unbounded; for $n > 10$, none.

This table does not allow any conclusions regarding the effect of the condition numbers of the linear systems on algorithm stability for large problems. We believe, however, that it gives some impression about how the number of iterations depends on the size of the problem. We can observe the same slow growth of the number of iterations with the dimensions m and n as has been observed throughout for interior-point methods for linear programs; see, e.g., [18], [7]. This makes interior-point methods particularly suitable for parallel computation, since the sequential part of the algorithm (the number of iterations) is small compared to the parallelizable part of solving a (large) system of linear equations at each iteration.

5.2. Boyd and Yang’s example. Here we use our method to solve two examples given in [3, Chap. 8]. Boyd and Yang modelled a simple two-input two-output control system and reduced the question of whether there exists a quadratic Lyapunov function that establishes stability of this control system to a problem of the form (1), involving 22 matrices $A^{(i)}$ of size 40×40 . In these examples, $A^{(0)} = 0$, and the problem was to find whether there exists a definite linear combination of the $A^{(i)}$, $i \geq 1$. For problem (1), there does, but problem (2) is bounded, i.e., $\lambda^{\text{opt}} = 0$ for $x = 0$. When solving these problems with our implementation, it takes one iteration, three evaluations of the Hessian, and twelve steps of (inexact) Newton’s method to find a negative definite combination for problem (1), and six iterations, nine Hessians, and 24 inexact Newton steps for problem (2) to discover that the optimal solution for problem (2) is within 10^{-4} of $5.7 \cdot 10^{-5}$. (It is exactly zero.) We emphasize that, also for problem (2), we can stop after *one* iteration, since, by Lemma 3, the existence of the first analytic center already implies boundedness of (1). A more careful imple-

³ This holds, with two exceptions where it was found in the second iteration. For the first iteration, we added the extra bound $\|x\| \leq 1000$ to avoid divergence for the unbounded problems.

TABLE 2

Problem (1)	$\lambda_{\max} < 0$	$\epsilon = 10^{-4}$	$\epsilon = 10^{-6}$
Interior point	5	8	11
Cutting plane	109	242	—
Subgradient	5507	—	—

TABLE 3

Problem (2)	$\epsilon = 10^{-4}$	$\epsilon = 10^{-6}$
Interior point	8	10
Cutting plane	113	—
Subgradient	—	—

mentation that determines only whether the first center exists or not will suffice for this type of problem and will be more efficient. (The case that there exists a nonzero positive semidefinite but not a positive definite combination will be hard to detect, however.) Boyd and Yang [3] solve a variation of problem (1)

$$\min_{|x_i| \leq 1} \lambda_{\max} \left(\sum_{i=1}^n x_i A^{(i)} \right).$$

This problem has a nonempty and finite set of optimal solutions for any set of symmetric $A^{(i)}$, while problem (1) does not. However, it is not well-posed in the sense that a definite combination of the two matrices $A^{(1)} := \text{diag}(1, -1, -1)$ and $A^{(2)} := \text{diag}(-1 + \epsilon, 1 + \epsilon, 1)$ might not be detected for very small $\epsilon > 0$. Boyd and Yang solve the problem by the cutting plane algorithm and by the subgradient method. In Tables 2 and 3, we list their results and those that we obtain from an interior-point method by imposing the same bounds on the problem as they. It is straightforward (but not efficient!) to modify the interior-point algorithm to this situation. The tables show the number of iterations necessary to determine that there is a negative combination or to find the optimal value up to an accuracy of 10^{-4} and of 10^{-6} . It should be noted that each iteration of the subgradient method involves an eigenvalue computation of a 40×40 matrix (this consists of five 8×8 blocks in this example, and is thus relatively cheap). In addition, the i th iteration of the cutting plane algorithm involves the solution of a linear program with $m + 1$ unknowns and $2m + i$ constraints. It is difficult to compare the computational effort of all three of these methods under consideration of the sparsity pattern of the matrices $A^{(i)}$, but, for $m < n$ and n sparse, as is the case here, we believe that an efficient implementation of the interior-point method may be very fast.

For problem (2), Boyd and Yang only determine that the problem has a largest eigenvalue that is larger than $-\epsilon$. We list the results for two different values of ϵ . Since the subgradient method does not provide a lower bound on the optimal solution, it fails for this example.

5.3. Overton's example. This is an example taken from Overton [25]. The problem is to find the diagonal of a 10×10 matrix M to minimize its maximum eigenvalue in absolute value. The off-diagonal elements of M are given. There are obvious modifications of the algorithm in this paper that do not increase the size of the problem and that take the "extreme sparsity" of the matrices $A^{(i)}$ —one nonzero

TABLE 4

iteration	λ_{\max}	rate of convergence
0	37.7	—
1	33.9	1.3
2	25.2	4.1
3	22.83	6.1
4	22.435	6.7
5	22.3763	6.8
6	22.3680	5.4
7	22.36642	6.2
8	22.366144	13.3

element on the diagonal for $1 \leq i \leq 10$ —into account. Here, however, we simply apply our algorithm with the matrices

$$\hat{A}^{(i)} := \begin{pmatrix} A^{(i)} & 0 \\ 0 & -A^{(i)} \end{pmatrix}, \quad 0 \leq i \leq 10,$$

thus doubling the size of the problem and working with 11 matrices of size 20. To test the limitations of our algorithm, we first deliberately set the stopping criterion to $1.0 \cdot 10^{-20}$. The optimal value λ^{opt} that our program computes for this problem is 22.36612164584166. This result is obtained in iteration 17 for $\mu = 1.14 \cdot 10^{-15}$ and $\rho = 1.18$, and it is accurate to at least 13 digits. The program stops in the next iteration with the Matlab message that the matrix $(A(\tilde{x}))$ to be inverted has a condition number of 10^{17} , and our algorithm then terminates with an error message. In Table 4, we run the problem with stopping criterion $n\mu/\rho \leq 10^{-4}$ and list the iteration number, λ_{\max} in that iteration, the value of μ , and the rates by which $\lambda_{\max} - \lambda^{\text{opt}}$ is reduced compared to the previous iteration. Note that the final accuracy is higher than the four digits guaranteed by Lemma 4. In these iterations, the Hessian is recomputed ten times, and a total of 25 (inexact) Newton steps are performed. Overton [25] presents an algorithm for solving this problem based on successive quadratic programming (QP). In each iteration, his algorithm solves a small number of quadratic programs with $m + 1$ unknowns and several equality constraints, the number depending on the estimated multiplicity of λ_{\max} at the solution. Overton also presents numerical experiments for the above problem and notes that “this problem is quite difficult to solve, since at the optimal solution the interior eigenvalues are nearly equal to λ_{\max} .” His program computes the optimal solution with ten digits of accuracy in 14 iterations solving a total of 26 QP’s. We test the interior-point method for different settings of the various parameter values, but we do not observe any irregular behaviour of the interior-point method. The observed robustness of the interior-point method coincides with the fact that the proof of convergence of the method is independent of the data and only depends on the dimensions m and n .

Overton [26] also suggests a successive linear programming algorithm for problem (1) that is better suited for large problems. Since our current implementation does not exploit sparsity, we cannot compare the two methods for large problems. (For small problems, Overton favours his SQP algorithm.) We note that both of Overton’s methods compute “dual matrix” information that is relevant for sensitivity analysis of the solution; see [26], as well as Appendix B and the proof of Lemma 4 in Appendix A.

We also apply the software package of Nesterov and Nemirovsky to solve this problem. Using the default option of short primal steps, their algorithm takes 23

iterations to obtain the solution 22.36614 (six digits of accuracy), and, using the option of large primal steps, the algorithm takes 15 iterations to find the solution 22.366124 (seven digits of accuracy). In both cases, the algorithm stops at this point with the error message “unable to compute the LU factorization.” We also test a number of test problems that are provided with the package. In all cases, the algorithm with large primal steps yield a slightly more accurate solution in less iterations, but, also, in all cases, the algorithm terminates with an error message after obtaining five to nine digits of accuracy. We believe that the linear algebra used by Matlab for our program is more stable than the one used by Nesterov and Nemirovsky, and this may be the most important reason for the observed difference in accuracy of the two programs. However, Nesterov and Nemirovsky do not use a centering step, and (as motivated in §4.6) this may be another reason for the higher accuracy of our algorithm⁴. For this problem, both programs (the one from Nesterov and Nemirovsky, as well as our Matlab version) are run with double precision on a 386 IBM PC.

It is hard to compare the computational effort of the algorithm of Nesterov and Nemirovsky with ours. For both algorithms, the computational effort in each iteration is dominated by the computation of the Hessian. Our algorithm uses a slightly smaller number of recomputations of the Hessian, but a larger number of additional operations per Hessian.

Appendix A. We briefly outline the proofs of Lemmas 1, 3, and 4.

Proof of Lemma 1. It is sufficient to prove that the function $\phi(x) := -\log \det_+(A(x))$ is strictly convex for all x for which $A(x)$ is positive definite. By the results of §4.3, the ij th entry of the Hessian of ϕ is given by

$$H_{ij} = BA^{(i)} \bullet A^{(j)}B.$$

Let $z \in \mathbb{R}^n$ be nonzero; then

$$z^T H z = B \left(\sum z_i A^{(i)} \right) \bullet \left(\sum z_j A^{(j)} \right) B = \left\| \left(\sum z_i A^{(i)} \right) B \right\|_F^2$$

is the square of the Frobenius norm of the nonzero matrix $(\sum z_i A^{(i)})B$ and thus positive. (The matrix is nonzero, since, by assumption, the $A^{(i)}$ are linearly independent and B is invertible.) \square

Proof of Lemma 3. 1) Suppose that λ^{opt} is finite, i.e., $\lambda^{\text{opt}} \geq 0$ by assumption on $A^{(0)}$. We show that the barrier function $\varphi(\cdot, \mu)$ has unique minimizers. By definition of \tilde{w} , the point \tilde{x}^0 is a unique minimum of $\varphi(\cdot, 1)$. (Its gradient is zero at \tilde{x}^0 , and, by Lemma 1, $\varphi(\cdot, 1)$ is strictly convex.) The existence of a unique minimum for the convex function $\varphi(\cdot, 1)$ implies that $\varphi(\tilde{x}, 1) \rightarrow \infty$ as $\tilde{x} \rightarrow \infty$. Now let $\mu \in (0, 1)$. Since $\lambda^{\text{opt}} \geq 0$, it follows that $\varphi(\tilde{x}, \mu) = \infty$ if $\lambda \leq 0$. For $\lambda > 0$, it follows trivially that $\varphi(\tilde{x}, \mu) > \varphi(\tilde{x}, 1)$; hence $\varphi(\tilde{x}, \mu) \rightarrow \infty$ as $\tilde{x} \rightarrow \infty$. Thus, the existence of a unique minimum for $\varphi(\cdot, \mu)$ follows again from strict convexity of $\varphi(\cdot, \mu)$.

2) Conversely, if $\lambda^{\text{opt}} = -\infty$ then there exists an $\bar{x} \in \mathbb{R}^m$ such that $\lambda_{\max}(\sum \bar{x}_i A^{(i)}) < -1$. For $\bar{x} := (\bar{x}, -1)^T$ and $\mu \in (0, 1)$ such that $\rho/\mu > \tilde{w}^T \bar{x}$, we find that $\varphi(t\bar{x}, \mu) \rightarrow -\infty$ as $t \rightarrow \infty$; i.e., a perturbed center does not exist for this μ .

3) Suppose that the analytic center exists for some $\mu > 0$. Then there is no $x \neq 0$ such that $\sum x_i A^{(i)} \leq 0$. (The $A^{(i)}$ are linearly independent, and $\varphi(\cdot, \mu)$ decreases

⁴ A modification of our algorithm that does without centering gave only five digits accuracy in 23 iterations.

along the ray given by $\tilde{x} = (x^T, 0)^T$ if $\sum x_i A^{(i)} \leq 0$.) Hence, there is an $\epsilon > 0$ such that $\lambda_{\max}(\sum x_i A^{(i)}) \geq \epsilon$ for all x with $\|x\| = 1$. Also, λ^{opt} is finite. For $\|x\| \geq 1 + \lambda^{\text{opt}}/\epsilon$, it follows that $\lambda_{\max}(A(x)) = \lambda_{\max}(A^{(0)} + \sum x_i A^{(i)}) \geq \lambda_{\min}(A^{(0)}) + \lambda_{\max}(\sum x_i A^{(i)}) \geq 0 + \epsilon(1 + \lambda^{\text{opt}}/\epsilon) = \lambda^{\text{opt}} + \epsilon$. Therefore, the compact domain $\|x\| \leq 1 + \lambda^{\text{opt}}/\epsilon$ contains all optimal solutions, and the set of optimal solutions is not empty.

4) Suppose that the set of optimal solutions is finite and nonempty. Let \bar{x} be an optimal solution. Then there exists a circle of radius r that contains the optimal set $S_{\text{opt}} \subset \bar{x} + K_r$ and $\lambda_{\max}(A(x)) \geq \lambda^{\text{opt}} + \epsilon$ for $x \in \partial(\bar{x} + K_r)$, the boundary of $\bar{x} + K_r$. Let $\|g\| = r$, then, by convexity, $\lambda_{\max}(A(\bar{x} + \theta g)) \geq \lambda^{\text{opt}} + \theta\epsilon$ for $\theta \geq 1$; hence, $\lambda > \lambda^{\text{opt}} + \theta\epsilon$ if $\lambda I - A(\bar{x} + \theta g) > 0$. Therefore, for any fixed $\mu > 0$, the quantity λ/μ grows at least linearly with θ , while $\log \det_+(A(\tilde{x}))$ is sublinear in θ , (the determinant is a multinomial in the matrix coefficients). For $\tilde{w} = 0$, this implies that $\varphi(\tilde{x}, \mu) \rightarrow \infty$ as $\tilde{x} \rightarrow \infty$, i.e., $\varphi(\cdot, \mu)$ has unique minimizers, i.e., the analytic centers exist for all $\mu > 0$. \square

Proof of Lemma 4. This lemma can either be shown directly using the results in [11] or more elegantly using the duality results of [1]. We show the second approach. By the results on Newton's method in §3.3, the assumption on \tilde{x} implies that the analytic center exists, and therefore, by Lemma 3, the set of optimal solutions of (1) is nonempty and bounded. First, suppose that \tilde{x} is on the central path, $\tilde{x} = \tilde{x}(\mu)$ for some $\mu > 0$ and $\tilde{w} = 0$. Then, by definition, the partial derivatives of φ are all zero, that is (by (17) and (18)),

$$(21) \quad \frac{\partial}{\partial \tilde{x}_i} \varphi(\tilde{x}, \mu) = A^{(i)} \bullet B(\tilde{x}) = 0 \quad \text{for } i = 1, \dots, m$$

and

$$(22) \quad \frac{\partial}{\partial \lambda} \varphi(\tilde{x}, \mu) = \frac{\rho}{\mu} - \text{trace} B(\tilde{x}) = 0.$$

Also from Appendix B, the dual problem as given in [1] is

$$\max_{Y \geq 0} \{A^{(0)} \bullet Y \mid \text{trace} Y = 1, A^{(i)} \bullet Y = 0\}.$$

Hence, the matrix $(\mu/\rho)B(\tilde{x})$ is dual feasible, and the “duality gap” can easily be computed as

$$\lambda - \frac{\mu}{\rho} B(\tilde{x}) \bullet A^{(0)} = -\frac{\mu}{\rho} \sum_{i=1}^{m+1} \tilde{x}_i (A^{(i)} \bullet B(\tilde{x})) - \frac{\mu}{\rho} B(\tilde{x}) \bullet A^{(0)}$$

(by substituting (21) and (22) and using $A^{(m+1)} = -I$)

$$= -\frac{\mu}{\rho} \left(A^{(0)} + \sum_{i=1}^{m+1} \tilde{x}_i A^{(i)} \right) \bullet B(\tilde{x}) = \frac{\mu}{\rho} A(\tilde{x}) \bullet B(\tilde{x}) = \frac{n\mu}{\rho},$$

and is an upper bound for $\lambda - \lambda^{\text{opt}}$. A similar criterion also holds for points \tilde{x} near the central path: Suppose that the Newton step $\Delta\tilde{x}$ starting at \tilde{x} for finding the analytic center $\tilde{x}(\mu)$ has H -norm less than $\sqrt{\alpha}/4$, $\|\Delta\tilde{x}\|_H \leq \sqrt{\alpha}/4$.

This assumption is equivalent to $\|\tilde{w} + g(\tilde{x}, \mu)\|_{H^{-1}} \leq \sqrt{\alpha}/4$, where $g(\tilde{x}, \mu)$ is the norm of the gradient of the perturbed barrier function defined in (18) for $\tilde{w} \neq 0$. Often (for example, if the optimal solution x^{opt} is unique), we can show that

$\lim_{\mu \rightarrow 0} \|\tilde{w}\|_{H(x(\mu))^{-1}} = 0$, so that the above-mentioned stopping criterion is also useful when following a path of perturbed centers.

By the inner ellipsoidal approximation of the feasible set in [23], [11], it follows in a straightforward manner that the above bound on $\lambda - \lambda^{\text{opt}}$ still holds in the weaker form stated in Lemma 4. \square

Proof of the remark in step 3) §4.4. We show that the last component of $D\varphi(\tilde{x} - \beta\tilde{x}', \mu)$ is a decreasing function of β for small $|\beta|$. Clearly,

$$\frac{d}{d\beta} D\varphi(\tilde{x} - \beta\tilde{x}', \mu)e_{m+1} = -\tilde{x}'^T H(\tilde{x} - \beta\tilde{x}')e_{m+1} = -e_{m+1}^T H^{-1}(\tilde{x})H(\tilde{x} - \beta\tilde{x}')e_{m+1}$$

is negative (and equal to -1) for $\beta = 0$. (Here, $H(\tilde{x}) = D^2\varphi(\tilde{x}, \mu)$ is independent of μ as noted in (19).) From the definition of self-concordance, we can show that this is also negative if $\|\beta\tilde{x}'\|_{H(\tilde{x})} \leq \sqrt{\alpha}/4$. \square

We note that (for the case of the barrier function of a linearly constrained problem) there are (very “artificial”) examples that $\hat{\mu}_{k+1} > \mu_k$ for large $\beta < \beta_{\max}$ even when the extrapolation started on the central path and $\hat{\mu}_{k+1} < \mu_k$ for smaller values of $\beta > 0$. In general, however, it is true that $\lim_{\beta \rightarrow \beta_{\max}} \hat{\mu}_{k+1} = 0$, and in our program the safeguard $\mu_{k+1} \leq \mu_k/2$ is never needed.

Appendix B. Duality. We briefly state a nice duality theorem for problem (5) presented by Alizadeh in [1].

A slightly more general form of problem (5), in that it allows a general linear objective function and does not fix $A^{(m+1)} = -I$, is the problem

$$\max_{x \in \mathbb{R}^{m+1}} \left\{ b^T x \mid C - \sum_i x_i A^{(i)} \geq 0 \right\}, \quad (\text{primal}),$$

where $C = A^{(0)}$. Alizadeh gives a short and self-contained proof that for this problem there is a dual problem

$$\min_{Y \geq 0} \{ C \bullet Y \mid A^{(i)} \bullet Y = b_i \}, \quad (\text{dual})$$

for which the following duality relations hold (the dual variable is a positive semidefinite matrix Y). Optimal solutions x^{opt} and Y^{opt} exist and the optimal values are the same if both problems have strictly feasible solutions. Furthermore, $b^T x \leq C \bullet Y$ for any primal and dual feasible variables x and Y . Alizadeh notes a surprising structural similarity of these problems to linear programming, develops a potential reduction method for the dual problem, and applies it to approximate the solution of combinatorial optimization problems.

The above duality result implies that the dual of problem (5) is given by

$$\max_{Y \geq 0} \{ A^{(0)} \bullet Y \mid Y \bullet I = 1, A^{(i)} \bullet Y = 0 \}.$$

Acknowledgments. The author thanks Prof. Michael Saunders for many helpful comments and for his warm hospitality at Stanford University. The author particularly thanks Prof. Stephen Boyd for bringing the problem and its applications to his attention and for many helpful discussions and suggestions, as well as for the data for the test problem in §5.2. Finally, the author thanks the referees for their constructive criticism.

REFERENCES

- [1] F. ALIZADEH, *Optimization over the positive definite cone: Interior-point methods and combinatorial applications*, in *Advances in Optimization and Parallel Computing*, P. Pardalos, ed., North-Holland, Amsterdam, 1992, pp. 1–25.
- [2] S. P. BOYD AND C. H. BARRAT, *Linear controller design: Limits of performance*, Prentice-Hall Information and System Sciences Series, T. Kailath, ed., Englewood Cliffs, NJ, 1990.
- [3] S. P. BOYD AND Q. YANG, *Structured and simultaneous Lyapunov functions for system stability problems*, *Internat. J. Control*, 49 (1989), pp. 2215–2240.
- [4] A. V. FIACCO AND G. P. MCCORMICK, *Sequential unconstrained minimization techniques*, in *Nonlinear Programming*, John Wiley, New York, 1968.
- [5] R. FLETCHER, *Semi-definite constraints in optimization*, *SIAM J. Control Optim.*, 23 (1985), pp. 493–513.
- [6] ———, *A new variational result for quasi-newton formulae*, *SIAM J. Optim.*, 1 (1991), pp. 18–21.
- [7] P. E. GILL, W. MURRAY, D. B. PONCELEON, AND M. A. SAUNDERS, *Preconditioners for indefinite systems arising in optimization*, Report SOL 90-8, Dept. of Operations Research, Stanford University, Stanford, CA, 1990.
- [8] C. J. GOH AND K. L. TEO, *On minimax eigenvalue problems via constrained optimization*, *J. Optim. Theory Appl.*, 57 (1988) pp. 59–68.
- [9] F. JARRE, *On the convergence of the method of analytic centers when applied to convex quadratic programs*, Report No. 35 (1987), Schwerpunktprogramm der DFG Anwendungsbezogene Optimierung und Steuerung; *Math. Programming*, 49 (1991), pp. 341–358. (In revised form.)
- [10] ———, *The method of analytic centers for solving smooth convex programs*, in *Lecture Notes in Mathematics*, Vol. 1405, S. Dolecki, ed., *Optimization, Proceedings Varetz 1988*, Springer, New York, 1989, pp. 69–85.
- [11] ———, *Interior-point methods for convex programming*, Report SOL 90-16, Dept. of Operations Research, Stanford University, Stanford, CA, 1990; *Appl. Math. Optim.*, to appear.
- [12] F. JARRE AND M. A. SAUNDERS, *An implementation of an interior-point method for convex programming*, Report SOL 91-9, Dept. of Operations Research, Stanford University, Stanford, CA, 1991.
- [13] F. JARRE, G. SONNEVEND, AND J. STOER, *An implementation of the method of analytic centers*, in *Lecture Notes in Control and Information Sciences 111, Analysis and Optimization of Systems*, A. Benoussan and J. L. Lions, eds., Springer, New York (1988) pp. 297–307.
- [14] V. A. KAMENETSKII, *Absolute stability and absolute instability of control systems with several nonlinear nonstationary elements*, *Automat. Remote Control*, 12 (1983), pp. 1543–1552.
- [15] V. A. KAMENETSKII AND E. S. PYATNITSKII, *Gradient method of constructing Lyapunov functions in problems of absolute stability*, *Automat. Remote Control*, 1 (1987), pp. 1–9.
- [16] N. KARMARKAR, *A new polynomial-time algorithm for linear programming*, *Combinator.*, 4 (1984), pp. 373–395.
- [17] C. LEMARECHAL AND R. MIFFLIN, EDS., *Nonsmooth Optimization*, Pergamon Press, Oxford, (1978).
- [18] I. J. LUSTIG, R. E. MARSTEN, AND D. F. SHANNO, *On implementing Mehrotra's predictor-corrector interior-point method for linear programming*, Report SOR 90-03, Dept. of Civil Engrg. and Oper. Res., Princeton University, Princeton, NJ, 1990.
- [19] S. MEHROTRA AND J. SUN, *An interior point algorithm for solving smooth convex programs based on Newton's method*, Report 88-08, Dept. of Industrial Engrg. and Management Sciences, Northwestern University, Evanston, IL, 1988.
- [20] C. MOLER, J. LITTLE, AND S. BANGERT, *PRO-MATLAB User's Guide*, The Math-Works, Inc., Sherborn, MA, 1987.
- [21] W. MURRAY AND M. L. OVERTON, *A projected Lagrangian algorithm for nonlinear minimax optimization*, *SIAM J. Sci. Statist. Comput.*, 1 (1980), pp. 345–370.
- [22] J. E. NESTEROV AND A. S. NEMIROVSKY, *A general approach to polynomial-time algorithms design for convex programming*, Report, Central Economical and Mathematical Institute, USSR Acad. Sci., Moscow, 1988.
- [23] ———, *Self-concordant functions and polynomial-time methods in convex programming*, Report, Central Economical and Mathematical Institute, USSR Acad. Sci., Moscow, 1989.
- [24] ———, *Optimization over positive semidefinite matrices*, Manual, Central Economical and Mathematical Institute, USSR Acad. Sci., Moscow, 1990.
- [25] M. L. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, *SIAM J. Matrix Anal. Appl.*, 9 (1988), pp. 256–268.

- [26] ———, *Large-scale optimization of eigenvalues*, NYU Computer Science Dept. Report No. 505, Courant Institute of Mathematical Sciences, New York University, New York, 1990.
- [27] M. L. OVERTON AND R. S. WOMERSLEY, *On minimizing the spectral radius of a nonsymmetric matrix function—Optimality conditions and duality theory*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 473–498.
- [28] E. R. PANIER, *On the need for special purpose algorithms for minimax eigenvalue problems*, Report, Dept. of Elec. Eng., University of Maryland, College Park, MD, 1989.
- [29] E. S. PYATNITSKII AND V. I. SKORODINSKII, *Numerical methods of construction of Lyapunov functions and absolute stability criteria in the form of numerical procedures*, Automat. Remote Control, 11 (1983), pp. 1427–1437.
- [30] U. T. RINGERTZ, *Optimal design of nonlinear shell structures*, Report, Flygtekniska Försöksanstalten, The Aeronautical Research Institute of Sweden, 1991.
- [31] H. SCHRAMM, *Eine Kombination von Bundle und Trust-Region-Verfahren zur Lösung nicht-differenzierbarer Optimierungsprobleme*, Ph.D. thesis, Universität Bayreuth, 1989.
- [32] G. SONNEVEND, *An 'analytical centre' for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming*, in Lecture Notes in Control and Information Sciences 84, System Modelling and Optimization, 12th IFIP Conference on Systems Optimization, 1985, Springer-Verlag, Berlin, New York, 1986, pp. 866–878.
- [33] ———, *A new method for solving a set of linear (convex) inequalities and its applications*, 5-th IFAC-IFORS Conference, Budapest 1986, Pergamon press (1987).
- [34] ———, *Applications of analytic centers to feedback control systems*, in Control of Uncertain Systems, D. Hinrichsen and B. Martensson, eds., Bremen, June 1989, Birkhäuser, Basel, Boston, (1990).
- [35] G. SONNEVEND AND J. STOER, *Global ellipsoidal approximations and homotopy methods for solving convex analytic programs*, Appl. Math. Optim., 21 (1989), pp. 139–166.
- [36] G. W. STEWART AND J. G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
- [37] Y. YE, O. GÜLER, R. A. TAPIA, AND Y. ZHANG, *A quadratically convergent $O(\sqrt{n}L)$ -iteration algorithm for linear programming*, Report 91-26, Dept. of Management Sciences, The University of Iowa, Iowa City, 1991.
- [38] J. ZOWE, *The BT-Algorithm for Minimizing a Nonsmooth Functional Subject to Linear Constraints*, manuscript, Universität Bayreuth, 1989.

CONDITIONS FOR OPTIMALITY OVER H^∞ *

J. WILLIAM HELTON[†] AND ORLANDO MERINO[‡]

Abstract. The fundamental optimization problem in worst-case frequency domain design where stability is the key constraint is

$$\begin{aligned} &\text{Given } \Gamma \text{ a map from } \mathbb{T} \times \mathbb{C}^N \text{ to } \mathbb{R}^+ \\ &\text{FIND } \gamma^* > 0 \text{ and } f^* \in A_N \text{ such that} \\ (\text{OPT}_{A_N}) \quad &\gamma^* = \inf_{f \in A_N} \sup_{\theta} \Gamma(e^{i\theta}, f(e^{i\theta})) = \sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta})). \end{aligned}$$

Here A_N denotes the \mathbb{C}^N -valued functions on the unit circle \mathbb{T} with analytic continuation to the unit disk \mathbb{D} that are continuous on the closed unit disk. The special case where the sublevel sets of Γ in z are “disks” in \mathbb{C}^N is the main mathematical problem in the area called H^∞ -control (cf. [B. A. Francis, *Lecture Notes in Control and Information Sci.*, Vol. 88, Springer-Verlag, Berlin, New York, 1986]), which has been one of the main emphases of control since the early 1980s. This article gives necessary conditions for f^* in A_N to be a solution to such a problem. These conditions are practical for computer implementation and are sufficient as well as necessary when $N \leq 2$ or when Γ is convex in the second variable. In §3, for $N = 2$, it is proved that there are very nice Γ (strictly pseudoconvex) producing an OPT problem having nonunique solutions. In §4 it is shown that coordinate descent algorithms often suggested by engineers for $N > 1$ sometimes do not attain the optimum (in practice, almost never). Section 5 treats another type of problem. Here $\tilde{\Gamma} \geq 0$ maps $\mathbb{T} \times \mathbb{R}^K \times \mathbb{C}^N$ to \mathbb{R} , and it is the goal to solve

$$(\text{UNCOPT-S}) \quad \inf_{f \in A_N} \sup_{\theta} \sup_{\alpha \in \mathcal{R}_{e^{i\theta}}} \tilde{\Gamma}(e^{i\theta}, \alpha, f(e^{i\theta})).$$

Here $\mathcal{R}_{e^{i\theta}}$ is a given closed set in \mathbb{R}^K . This problem is basic to engineering situations where there is uncertainty in the accuracy of mathematical models.

Key words. minimax optimization, sup-norm optimization, analytic functions

AMS subject classifications. 49K35, 49J35, 32A35, 30D55, 93B36

1. Introduction. The fundamental optimization problem in worst-case frequency domain design where stability is the key constraint is

$$\begin{aligned} &\text{Given } \Gamma \text{ a map from } \mathbb{T} \times \mathbb{C}^N \text{ to } \mathbb{R}^+, \\ &\text{FIND } \gamma^* > 0 \text{ and } f^* \in A_N \text{ such that} \\ (\text{OPT}_{A_N}) \quad &\gamma^* = \inf_{f \in A_N} \sup_{\theta} \Gamma(e^{i\theta}, f(e^{i\theta})) = \sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta})). \end{aligned}$$

Here A_N denotes the \mathbb{C}^N -valued functions on the unit circle \mathbb{T} with analytic continuation to the unit disk \mathbb{D} , which are continuous on the closed unit disk. (See [Fr], [H1], and [H2].) The special case where the sublevel sets

$$S_\theta(c) = \{z \in \mathbb{C}^N : \Gamma(e^{i\theta}, z) \leq c\}$$

of Γ in z are “disks” in \mathbb{C}^N is the main mathematical problem in the area called H^∞ -control (cf. [Fr]), which has been one of the main emphases of control since the early 1980s. This article gives necessary conditions for f^* in A_N to be a solution to such a problem. These conditions are practical for computer implementation and are sufficient as well as necessary when $N \leq 2$ or when Γ is convex in the second variable. In §3 we prove that for $N = 2$ there are very nice Γ (strictly pseudoconvex) producing an OPT problem having nonunique

* Received by the editors February 12, 1990; accepted for publication (in revised form) March 12, 1992. This work was supported in part by the Air Force Office of Scientific Research and the National Science Foundation.

[†] Department of Mathematics, University of California, San Diego, La Jolla, California 92093.

[‡] Department of Mathematics, Texas Tech University, Lubbock, Texas 79409.

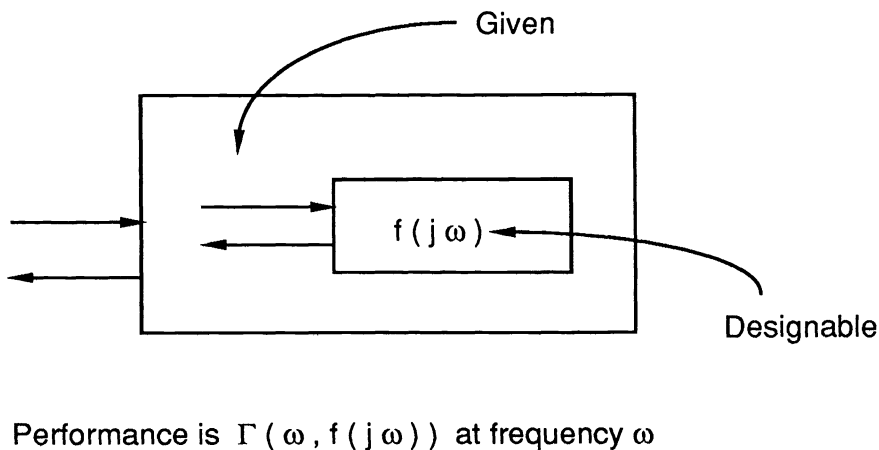


FIG. 1.1

solutions. In §4 we show that coordinate descent algorithms often suggested by engineers for $N > 1$ sometimes do not attain the optimum (in practice, almost never). Section 5 treats another type of problem. Here $\tilde{\Gamma} \geq 0$ maps $\mathbb{T} \times \mathbb{R}^K \times \mathbb{C}^N$ to \mathbb{R} , and we wish to solve

$$(\text{UNCOPT-S}) \quad \inf_{f \in A_N} \sup_{\theta} \sup_{\alpha \in \mathcal{R}_{e^{i\theta}}} \tilde{\Gamma}(e^{i\theta}, \alpha, f(e^{i\theta})).$$

Here $\mathcal{R}_{e^{i\theta}}$ is a given closed set in \mathbb{R}^K . This problem is basic to engineering situations where there is uncertainty in the accuracy of mathematical models, for example, in the area of robust control [Dor], [Doy], [H1].

The problem UNCOPT is strictly more complicated than OPT and indeed has OPT as one “pure component” of it. To see this, for given $\tilde{\Gamma}$, define Γ by “solving”

$$(\text{UNC}) \quad \Gamma(e^{i\theta}, z) \triangleq \sup_{\alpha \in \mathcal{R}_{e^{i\theta}}} \tilde{\Gamma}(e^{i\theta}, \alpha, z).$$

Then, solving UNCOPT is equivalent to solving UNC and OPT for Γ . (Refer to our computer program for solving OPT. The results of §2 are used to compute diagnostics printed out at each iteration.)

1.1. Motivation. The OPT problem is central to the design of a system where specifications are given in the frequency domain and stability is a key issue. Suppose that our objective is to design a system, part of which we are forced to use (in control, it is called the plant) and part of which is designable; see Fig. 1.1. The objective of the design is to find the admissible f that gives the best performance. If we denominate performance Γ as a “cost,” then the big Γ is bad, and the “worst case” is the frequency ω at which

$$\sup_{\omega} \Gamma(\omega, f(j\omega))$$

occurs. We seek to minimize this over all admissible f . The stipulation that the designable part of the circuit be stable amounts to requiring that f has no poles in the right half plane (R.H.P.). In other words, $f \in A_N$ (R.H.P.). This is exactly the R.H.P.-version OPT problem. Even when parts of the system other than the designable part are in H^∞ , we can frequently reparametrize to get OPT. Consequently, OPT arises in a large class of problems.

Indeed, the OPT problem is so basic it might be called *the fundamental H^∞ problem of control*. This sits in distinction to the fundamental problem of H^∞ control, and there are varying opinions as to what that is.

Graphic interpretations of OPT are informative and useful. A sublevel set $\mathcal{L}_\omega(c) = \{z \in \mathbb{C}^N : \Gamma(\omega, z) < c\}$ of Γ is just the set of designable parameters z that give performance better than c . The objective is to find a function f with no poles in the R.H.P. so that each $f(j\omega)$ belongs to $\mathcal{L}_\omega(c)$. Any such f makes the performance of the overall system at least as good as c for all ω . Thus OPT corresponds to find such f for the smallest value of c for which f can be found. One example is the Horowitz templates of control; thus this article gives an optimality theory for a significant portion of Horowitz control.

Related frequency domain design theories are those originated by Mayne and Polak and more recently pursued by their students (e.g., Tits and Fan). In a sense, the problem they study is more general than OPT; though, technically speaking, it is not (since it is finite-dimensional). They optimize performance over classes of rational functions (which, of course, are highly nonconvex). We optimize over the class of all analytic functions on the R.H.P. and exploit their special structure to obtain the elegant and sharp optimality conditions found in this paper. Ironically, infinite-dimensionality of A_N helps because it guarantees that, at optimum, “every constraint is active.” (This is condition (i) of Theorem 2.2.) In another direction, linear programming is a viable approach to OPT problems, provided that Γ has sublevel sets $\mathcal{L}_\theta(c)$ that are convex. Then, we can approximate the $\mathcal{L}^\theta(c)$ by a simplex and use standard packages. This has been explored extensively by Boyd in [BB], and crude numerical comparisons of linear programming to other H^∞ methods were done in [BHM], [HS]. The theorems given have required that Γ be a smooth function; they do not apply as stated to a simplex in \mathbb{C}^N , since a simplex has corners. They would apply to a smooth approximation of the simplex, and so they might well be adaptable for use in linear-programming-based codes. Also, for $N = 1$, a general version of Theorem 2.1 appears in [HM] that applies to $\mathcal{L}_\theta(c)$, which are simplicies.

When each $\mathcal{L}_\omega(c)$ is a “disk” such as

$$S_\omega(c) = \left\{ (z_1, z_2, \dots, z_N) \in \mathbb{C}^N : \sum_{i=1}^N p_i(j\omega) |K_i(j\omega) - z|^2 \leq c \right\}$$

for some complex-valued functions K_i and positive-valued functions p_i in \mathbb{C}^N , the problem has an *explicit solution*. Other solvable OPT problems are easy to recognize, since so few of them exist. Crudely speaking, OPT_E with $E = H_n^\infty$ (see §1.2), is completely solved below:

- (i) when each $\mathcal{L}_\omega(c)$ is a disk on \mathbb{C}^N ,
- (ii) when each $\mathcal{L}_\omega(c)$ is a matrix ball,
- (iii) when each $\mathcal{L}_\omega(c)$ is a ball of symmetric matrices,
- (iv) when each $\mathcal{L}_\omega(c)$ is a ball of antisymmetric matrices.

A function Γ each of whose sublevel sets $\mathcal{L}_\omega(c)$ are disks in this sense will be called *quasi-circular*. (There is much literature on solutions to quasi-circular problems. See [BGR], [Dym], [FF], [Fr], [GM], [H4], and [Yng].)

Example 1. The famous “mixed sensitivity” performance measure of control is

$$(1.1) \quad \tilde{\Gamma}(\omega, T) \triangleq W_1(j\omega)|T - 1|^2 + W_2(j\omega)|T|^2,$$

where W_1 weights low frequencies and W_2 weights high frequencies. Interpolation constraints are easy to put in and convert $\tilde{\Gamma}$ to a Γ of a similar form. These problems are quasi-circular, as are the famous 1, 2, and 4 block problems of H^∞ control.

Example 2 (power mismatch and amplifier design). Classical control is very close to the work of Bode on amplifier design. Likewise, H^∞ -control is mathematically the same (actually a little easier) than the subject of gain equalization (or even noise suppression) in amplifiers. Historically, Youla and Saito [YS], gave a theory of single-input, single-output (SISO) gain equalization, and Helton gave a theory of multiple-input, multiple-output (MIMO) design (surveyed in [H5]), which preceded H^∞ -control and which contributed to its inception. These paradigm problems are quasi-circular.

Example 3. Two competing constraints typically yield \mathcal{L}_θ , which are intersections of two disks, and so forth.

Example 4. Bercovici, Foias, and Tannenbaum have recent results on Γ 's that are based on the spectral radius of a matrix. These are highly nonsmooth Γ 's and demonstrate interesting properties. This arises in a class of control problems with plant uncertainty.

Uncertainty in the mathematical model for a physical system, which, in control, is called plant uncertainty, naturally leads to OPT problems with very complicated Γ . Our formulation is to start with a performance measure $\tilde{\Gamma}$ that depends on what we believe the plant P to be at frequency ω and the choice T of the designable parameter at ω . The basic design optimization problem is

$$(\text{UNCOPT-S}_{\text{RHP}}) \quad \inf_{T \in A_N^\infty} \sup_{\omega} \sup_{p \in R_\omega} \tilde{\Gamma}(\omega, p, T(j\omega)).$$

Here R_ω denotes the range of values p at frequency ω that we believe the plant $P(j\omega)$ might actually take. This, of course, is the R.H.P. version of (UNCOPT-S). For this problem, “tightening the specs” amounts to calculating the “tightened” performance measure

$$(\text{UNC}_{\text{RHP}}) \quad \Gamma(\omega, T) = \sup_{p \in R_\omega} \tilde{\Gamma}(\omega, p, T).$$

After this is done, solving the full UNCOPT-S_{RHP} problem is equivalent to an OPT on the R.H.P. Thus OPT is the pure H^∞ part of UNCOPT (which is another good reason for calling OPT the fundamental H^∞ problem of control).

Plant uncertainty when treated in this way simply amounts to a mathematization of the age old engineering adage: “In the presence of uncertainty, tighten the specs.”

The maximization in UNC is time consuming and is a subject unto itself (the structured singular value (s.s.v.) and environs). Consequently, we certainly expect that the most effective numerical algorithms at the $(k + 1)$ th iteration will update current guess f^k and $P^k(j\omega)$ by doing a UNC step to increase $\tilde{\Gamma}(\omega, P, f^k(j\omega))$ for each ω , then an f step to decrease $\|\Gamma\|_\infty$, then another P step, then another f step, and so forth.

Doyle's μ -synthesis (after substantial reparameterization and a compromise) is such an algorithm. It is very natural, since the function $\Gamma(\omega, p, z)$ of UNCOPT is in many control applications for fixed p quasi-circular. Thus coordinate optimization produces a sequence $P^k \in L^\infty$ and $f^k \in H^\infty$, where the update for f^k comes by solving a quasi-circular OPT problem. We could imagine an infinite variety of choices between spending a long time on each P maximization step before going to an f step, or vice versa. Consequently, it is clear that, to begin a systematic study of the H^∞ plant uncertainty problem UNCOPT',

we should analyze two extreme situations. The *first* is where we do the p maximization completely; this is UNC. It is in principle studied by the s.s.v. school. The *second* is OPT for the “tightened” performance measure Γ . (For numerous articles on uncertainty and robustness, see [Dor].)

1.2. Notation. For convenience, we refer to OPT_E as the problem OPT, where the minimization is over the set $E \subset H_N^\infty$, and f^* is in E . By L_N^∞ , we represent the set of all \mathbb{C}^N -valued, essentially bounded measurable functions on \mathbb{T} , and H_N^∞ consists of those elements of L_N^∞ that extend to bounded analytic functions on \mathbb{D} . The space H_N^1 is defined similarly as consisting of \mathbb{C}^N -valued measurable functions F on \mathbb{T} that extend analytically to \mathbb{D} and have finite L^1 -norm, $\|F\|_{L^1} = \int_0^{2\pi} \|F(e^{i\theta})\|_{\mathbb{C}^N} (d\theta/2\pi) < \infty$. If $E \subset H_N^\infty$, set

$$(1.1) \quad RE \triangleq \{f \in E : \overline{f(e^{-i\theta})} = f(e^{i\theta}) \text{ a.e. on } \mathbb{T}\}.$$

Another way of defining RE is as the set of elements of E with real Fourier coefficients. Frequency response functions in engineering are all real on the real axis. Thus spaces such as RH_N^∞ , RA_N , and the corresponding OPT problems are the only ones that occur in engineering applications.

We investigate local conditions for solutions to OPT. For this, we use the terms *optimizer* and *directional optimizer*, which we now define below.

DEFINITION 1.1. Let $E \subset H_N^\infty$ and $f^* \in E$. The function f^* is a *local optimizer* for OPT_E if there exists an open set $V \subset E$ such that $f^* \in V$ and

$$(1.2) \quad \sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta})) \leq \sup_{\theta} \Gamma(e^{i\theta}, f(e^{i\theta})) \quad \forall f \in V.$$

The function f^* is a *directional optimizer* for OPT_E if, for each $h \in E$, there exists $t_h > 0$ such that

$$(1.3) \quad \sup \Gamma(e^{i\theta}, f^*(e^{i\theta})) \leq \sup \Gamma(e^{i\theta}, f^*(e^{i\theta}) + t h(e^{i\theta})) \quad \forall t \in (0, t_h).$$

The terms *strict local optimizer* and *strict local directional optimizer* are defined similarly, with strict inequality in (1.2) and (1.3), respectively.

Of course, if E is finite-dimensional, the concepts local optimizer and directional optimizer coincide. This is a simple consequence of the compactness of the unit ball in E . In the infinite-dimensional case, a strict directional optimizer may not be a strict local optimizer.

Now we establish some notation needed throughout the paper. Elements $z \in \mathbb{C}^N$ are represented as column vectors. By z^\dagger , we denote the transpose of z , so that $\bar{z}^\dagger z = \|z\|_{\mathbb{C}^N}^2$, and, if z has entries $z_l, l = 1, \dots, N$, then $\text{Re } z$ represents the column vector with entries $\text{Re } z_l, l = 1, \dots, N$. Let $G : \mathbb{C}^N \rightarrow \mathbb{R}$ be a class C^3 function. Then the order 2 expansion of G about z_0 is

$$(1.4) \quad \begin{aligned} G(z_0 + w) = & G(z_0) + 2 \text{Re } \frac{\partial G}{\partial z}(z_0)^\dagger w \\ & + \frac{1}{2} \left\{ 2\bar{w}^\dagger \frac{\partial^2 G}{\partial \bar{z} \partial z}(z_0) w + 2 \text{Re } \left(w^\dagger \frac{\partial^2 G}{\partial z^2}(z_0) w \right) \right\} \\ & + \mathcal{O}(\|w\|_{\mathbb{C}^N}^3), \end{aligned}$$

where $\partial G / \partial z(z_0)$ is a column vector on entries $\partial G / \partial z_l(z_0), l = 1, \dots, N$ and $\partial^2 G / \partial \bar{z} \partial z(z_0)$ (respectively, $\partial^2 G / \partial z^2(z_0)$) is an $N \times N$ matrix, with entries $\partial^2 G / \partial \bar{z}_l \partial z_j(z_0)$ (respectively, $\partial^2 G / \partial z_l \partial z_j(z_0)$), $j, l = 1, \dots, N$.

2. Necessary conditions and sufficient conditions for solutions to OPT. The following theorem of Helton in [H3] gives a clean-cut characterization of functions $f \in A_1$ that solve the problem OPT.

THEOREM 2.1 ($N = 1$). *Let $\Gamma : \mathbb{T} \times \mathbb{C} \rightarrow \mathbb{R}^+$ be of class C^2 . A continuous function $f^* \in A_1$ for which $a(e^{i\theta}) = (\partial\Gamma/\partial z)(e^{i\theta}, f^*(e^{i\theta}))$ is never zero is a strict local directional optimizer if and only if (i) $\Gamma(e^{i\theta}, f^*(e^{i\theta}))$ is constant in θ , and (ii) the winding number of a about 0 is positive. \square*

See [A] for a definition of winding number of a curve. It is also known that, for $N = 1$, under fairly general hypotheses, the solution exists, is unique, and has certain degree of smoothness (see [HM]). Thus Theorem 2.1 gives a practical test when $N = 1$ for determining if a given function $f^* \in A_1$ solves OPT.

We see that the situation for the case where $N > 1$ is quite different. Uniqueness no longer holds—this is the subject of §3.

Also, conditions (i) and (ii) can be generalized to the case where $N > 1$, but they are not sufficient to guarantee that a particular function $f^* \in A_N$ is a solution to OPT_{A_N} .

2.1. Direction optima. We start the study of the general case ($N \geq 1$) by presenting a generalization of (i) and (ii) of Theorem 2.1. For a given function $a : \mathbb{T} \rightarrow \mathbb{C}^N$, we write

$$N(a) = \{h \in H_N^\infty : a^t(e^{i\theta})h(e^{i\theta}) = 0, \text{ a.e. on } \mathbb{T}\}.$$

THEOREM 2.2. *Let Γ be of class C^3 and left $f^* \in A_N$ be such that $(\partial\Gamma/\partial z)(e^{i\theta}, f^*(e^{i\theta}))$ never equals 0 on \mathbb{T} . If $f^* \in A$ is a local directional optimizer, then*

(I) *The function $\Gamma(\cdot, f^*(\cdot))$ is constant on \mathbb{T} ,*

(II) *There exist functions $F \in H_N^1$ and $\lambda : \mathbb{T} \rightarrow \mathbb{R}^+$ measurable and positive almost everywhere on \mathbb{T} such that*

$$\frac{\partial\Gamma}{\partial z}(e^{i\theta}, f^*(e^{i\theta})) = \lambda(e^{i\theta})\chi(e^{i\theta})F(e^{i\theta}) \quad \text{a.e. on } \mathbb{T}.$$

Here χ is the function $\chi(e^{i\theta}) = e^{i\theta}$;

(III) *For every $h \in N((\partial\Gamma/\partial z)(\cdot, f^*(\cdot))) \cap A_N$, $h \neq 0$,*

$$\sup_{\theta} \left\{ \overline{h(e^{i\theta})}^\dagger \frac{\partial^2\Gamma}{\partial z \partial \bar{z}}(e^{i\theta}, f^*(e^{i\theta}))h(e^{i\theta}) + \text{Re} \left(h(e^{i\theta})^\dagger \frac{\partial^2\Gamma}{\partial z^2}(e^{i\theta}, f^*(e^{i\theta}))h(e^{i\theta}) \right) \right\} \geq 0.$$

Conversely, conditions (I)–(III) with strict inequality imply that f^* is a strict local directional optimizer.

To see that Theorem 2.2 generalizes Theorem 2.1, we obtain the latter from the former, by directly comparing (i), (ii) of Theorem 2.1 with (I)–(III) of Theorem 2.2. This we do below.

Since (i) and (I) are identical, we now look at (ii) versus (II). If (ii) is assumed, then we can write $a(e^{i\theta}) = (e^{i\theta})^n a_1(e^{i\theta})$, where n is the winding number of a about 0. In particular, $\text{wind}(a_1; 0) = 0$, so we can write $a_1(e^{i\theta}) = |a(e^{i\theta})|e^{u(e^{i\theta})i}$, where u is continuous and real-valued. If u^* is the harmonic conjugate of u , then

$$a_1 = \lambda(e^{i\theta})e^{i\theta}F(e^{i\theta}),$$

where $F(e^{i\theta}) = (e^{i\theta})^{n-1}e^{-u^*+iu}$ is in H_1^1 and $\lambda(e^{i\theta}) = |a(e^{i\theta})| \cdot e^{u^*(e^{i\theta})}$ is positive measurable. Thus (II) holds. This can be reversed to show that (II) implies (ii).

We now claim that (III) is trivially satisfied when $N = 1$ and (II) holds. Indeed, if $h \in H_1^\infty$ satisfies $(\partial\Gamma/\partial z)(e^{i\theta}, f^*(e^{i\theta})) \cdot h(e^{i\theta}) = 0$ almost everywhere on \mathbb{T} , we must have $h = 0$, since F is not the function zero. Thus $N((\partial\Gamma/\partial z)(\cdot, f^*(\cdot))) = \{0\}$.

If the function Γ is *real symmetric*, i.e.,

$$(2.1) \quad \Gamma(e^{i\theta}, z) = \Gamma(e^{-i\theta}, \bar{z}) \quad \forall e^{i\theta} \in \Pi, \forall z \in \mathbb{C}^N,$$

then we can consider the problem OPT_{RA_N} and try to obtain a version of Theorem 2.2 specialized to this situation. We have the following (partial) result.

THEOREM 2.2'. *In addition to the hypotheses to Theorem 2.2, assume that $f \in RA_N$ and that Γ satisfies (2.1). Modify conditions (II) and (III) of Theorem 2.2 by requiring that $F \in RH_N^1$ in statement (II), and that $h \in RN$ $\partial\Gamma/\partial z(\cdot, f^*(\cdot))$.*

Then (I)–(III) are necessary for $f^ \in RA_N$ to be a local directional optimizer for OPT_{RA_N} .*

Given Γ and f^* as in Theorem 2.2, we see that we can associate to the function

$$(2.2) \quad a(\cdot) = \frac{\partial\Gamma}{\partial z}(\cdot, f^*(\cdot))$$

defined on \mathbb{T} and valued in \mathbb{C}^N , a unique positive integer $w(a)$ (see Theorem 2.5). When $N = 1$, the integer $w(a)$ is precisely the winding number about zero of the function a .

DEFINITION 2.3. Let $b \in L_N^\infty$. A function $F \in H_N^1$ is an *analytic direction* for b if $\|F\|_{H_N^1} = 1$ and if there exists a measurable function $\lambda: \mathbb{T} \rightarrow \mathbb{R}^+$ such that

$$b(e^{i\theta}) = \lambda(e^{i\theta})F(e^{i\theta}) \quad \text{a.e. on } \mathbb{T}.$$

Hence (II) of Theorem 2.2 is equivalent to stating that $\chi^{-1}(\cdot)(\partial\Gamma/\partial z)(\cdot, f^*(\cdot))$ has an analytic direction.

DEFINITION 2.4. Let $b \in L_N^\infty$, set $\omega(b) \triangleq \sup\{m \in \mathbb{Z} : \chi^{-m}b \text{ has an analytic direction}\}$.

THEOREM 2.5. *Let Γ be as in the hypotheses to Theorem 2.2. If $f^* \in A_N$ is an optimizer for OPT_E and if $a(\cdot) = \partial\Gamma/\partial z(\cdot, f^*(\cdot))$, then*

- (1) $1 \leq \omega(a) < \infty$,
- (2) $\chi^{-\omega(a)} a$ has a unique analytic direction F_0 . Moreover, F_0 is a strong outer-function,¹
- (3) Let $F \in H_N^1$ be such that $\|F\|_{L^1} = 1$. Then F is an analytic direction for a if and only if there exist a constant $\mu > 0$ and elements $z_1, \dots, z_{\omega(a)}$ in the closed unit disk in \mathbb{C} such that

$$(2.3) \quad F(e^{i\theta}) = \prod_{l=1}^{\omega(a)} (e^{i\theta} - z_l)(1 - \bar{z}_l e^{i\theta}) F_0(e^{i\theta}) \quad \text{a.e. on } \mathbb{T},$$

- (4) If $f \in RA_N$ and Γ satisfies (2.1), then the function F_0 in 2 is in RH_N^1 , and the zeros z_l in (2.3) come in conjugate pairs.

See Theorem A.3 in Appendix A.

It is shown in [H3] that, when a has rational entries $a_j, j = 1, \dots, N$, we can compute an integer L out of the zeros and poles of the a_j 's, so that a necessary condition for optimality is $L > 0$. The number L proves to be precisely $\omega(a)$. Hence, in this case, (II) amounts to saying $\omega(a) \geq 1$.

THEOREM 2.6. *Assume the hypotheses of Theorem 2.2 (or Theorem 2.8). Let Γ and f^* be as in Theorem 2.2 (or Theorem 2.8). We have that*

¹ F_0 is strong outer if F_0 is an outer function such that $(\chi - e^{i\varphi})^{-2}F_0 \notin H_N^1$, for every $e^{i\varphi} \in \mathbb{T}$.

(1) If the function a can be extended meromorphically to a neighborhood of \mathbb{D}^- , then $\omega(a) = z(a) - p(a)$, where by definition $z(a) =$ number of zeros in \mathbb{D} plus number of double zeros on $\partial\mathbb{D}$, common to all the a_j 's, counting multiplicities,

$p(a) =$ total number of poles in \mathbb{D} plus total number of double poles on $\partial\mathbb{D}$, of the a_j 's;

(2) If, for some j_0 , the coordinate function a_{j_0} is never 0 on \mathbb{T} and has winding number $\omega(a_{j_0})$ about 0, then $1 \leq \omega(a) \leq \omega(a_{j_0})$. If also $\omega(a_{j_0}) = 1$, then $a_j/a_{j_0} \in H_1^\infty$, for $j = 1, \dots, N$.

The necessity of (I) and $\omega(a) > 0$ was first demonstrated for $N \geq 1$ in [H3] under stronger hypotheses than in Theorem 2.2.

We have found that Theorem 2.6 is particularly valuable in practice. We have a code for solving iteratively OPT, and, after each iterate f^k is produced, tests corresponding to (I) and (II) are performed. When $N = 1$, numerical experiments (to be reported elsewhere) show that

$$e^k = \left(\sup_{\theta} \Gamma(e^{i\theta}, f^k(e^{i\theta})) - \inf_{\theta} \Gamma(e^{i\theta}, f^k(e^{i\theta})) \right) / \sup_{\theta} \left\| \frac{\partial \Gamma}{\partial z}(e^{i\theta}, f^k(e^{i\theta})) \right\|_{\mathbb{C}^N}$$

is a good predictor of the true error $\|f^k - f^*\|_\infty$, while the winding number $\omega(a^k)$, where $\omega(a^k) = \omega(\partial\Gamma/\partial z(\cdot, f^k))$, very quickly stabilizes and proves to be unimportant. However, when $N > 1$, the situation is quite different: e^k tends to zero at superlinear rate, but $\|f^k - f^*\|_\infty$ decreases at a much slower rate. Since we do not have a good estimate of $\|f^k - f^*\|_\infty$, we employ (2) of Theorem 2.6 to produce a diagnostic that works well in combination with e^k . It consists in first locating $a_{j_0}^k = \partial\Gamma/\partial z(\cdot, f^*)$, which is never zero on \mathbb{T} and has $w(a_{j_0}^k) = 1$. Then we compute that

$$r^k = \sum_{j=1}^N \|P_{H_1^{2\perp}}(a_j^k/a_{j_0}^k)\|_{L^2},$$

where $P_{H_1^{2\perp}}$ denotes the orthogonal projection onto $H^{2\perp}$. Note that $r^k = 0$ implies that (II) holds,² so, as $f^k \rightarrow f^*$ in a computer run, we observe that $f^k \rightarrow 0$.

Until this point, we have given versions of (I) and (II) that are useful for computation. Unfortunately, (III) remains unwieldy, since we must check a certain condition for all analytic functions. How to overcome this remains the main open question in this area of necessary and sufficient conditions for optimality. However, when $N = 2$, we obtain an elegant and practical theorem.

THEOREM 2.7. *Set $N = 2$ in Theorem 2.2 and let $a(\cdot) = \partial\Gamma/\partial z(\cdot, f^*(\cdot))$ and $b(\cdot) = (-\partial\Gamma/\partial z_2(\cdot, f^*(\cdot)), \partial\Gamma/\partial z_1(\cdot, f^*(\cdot)))^\dagger$. Furthermore, suppose that the function $\chi^{-\omega(a)}$. a has an analytic direction F_0 that is continuous and nonzero on \mathbb{T} . Then (III) is equivalent to the statement below.*

At least one of the following statements is true:

(i) *There exists $e^{i\theta} \in \mathbb{T}$ such that*

$$b^\dagger(e^{i\theta})A(e^{i\theta})b(e^{i\theta}) \geq |b^\dagger(e^{i\theta})B(e^{i\theta})b(e^{i\theta})|;$$

(ii) *$b^\dagger Bb$ is never zero on \mathbb{T} , and either*

(a) *$\omega(b^\dagger Bb) > 2\omega(a)$, or*

² Of course, such an $a_{j_0}^k$ (with no zeros on \mathbb{T} and winding number about 0 equal to 1) does not always occur. However, for generic Γ , [M1] very strongly suggests $a_{j_0}^k$ as above do occur, at least for f^k close to f^* .

(b) $\omega(b^\dagger Bb) < 2\omega(a)$, and $\omega(b^\dagger Bb)$ is odd.

We illustrate use of the theorems above with an example.

Example 1. For $N = 2$ and $\varepsilon > 0$, let

$$\Gamma_\varepsilon(e^{i\theta}, z_1, z_2) = |100 + z_1 e^{i\theta} + 0.1(z_1 z_2 + z_1 + z_2)|^2 \\ + |100 + z_2 e^{i\theta} + 0.1(z_1 z_2 + z_1 + z_2)|^2 + \varepsilon |z_1|^2 + \varepsilon |z_2|^2.$$

We have that

$$\frac{\partial \Gamma_\varepsilon}{\partial z_1}(e^{i\theta}, z_1, z_2) = \overline{(100 + z_1 e^{i\theta} + 0.1(z_1 z_2 + z_1 + z_2))} \cdot (e^{i\theta} + 0.1 + 0.1 z_2) \\ + \overline{(100 + z_2 e^{i\theta} + 0.1(z_1 z_2 + z_1 + z_2))} \cdot (0.1 + 0.1 z_2) + \varepsilon \overline{z_1}$$

and

$$\frac{\partial \Gamma_\varepsilon}{\partial z_2}(e^{i\theta}, z_1, z_2) = \overline{(100 + z_1 0.1(z_1 z_2 + z_1 + z_2))} (0.1 + 0.1 z_1) \\ + \overline{(100 + z_2 e^{i\theta} + 0.1(z_1 z_2 + z_1 + z_2))} (e^{i\theta} + 0.1 + 0.1 z_1) + \varepsilon \overline{z_2}.$$

We now show that the constant function $f(e^{i\theta}) = (0, 0)$ satisfies (I) and (II), but is not an optimizer for Γ_ε , if $\varepsilon < 19$.

Now (I) holds for

$$\Gamma_\varepsilon(e^{i\theta}, 0, 0) = 20,000 \quad \forall e^{i\theta} \in \mathbb{T}.$$

To check (II), evaluate the partial derivatives of Γ_ε to obtain

$$\left(\frac{\partial \Gamma_\varepsilon}{\partial z_1}(e^{i\theta}, 0, 0), \frac{\partial \Gamma_\varepsilon}{\partial z_2}(e^{i\theta}, 0, 0) \right) = (100e^{i\theta} + 20, 100e^{i\theta} + 20).$$

Note that $100e^{i\theta} + 20 = 100|e^{i\theta} + 0.2|^2(e^{i\theta}/(1 + 0.2e^{i\theta}))$, so that (II) holds.

We claim that $f^*(0, 0)$ does not solve OPT when $\varepsilon < 19$. Begin by computing $b^\dagger Ab$ and $b^\dagger Bb$ at $(0, 0)$, as follows:

$$\overline{b}^\dagger Ab|_{z_1=0, z_2=0} \\ = \begin{pmatrix} -100e^{i\theta} - 20 \\ 100e^{i\theta} + 20 \end{pmatrix}^\dagger \begin{pmatrix} 1.02 + 0.1e^{i\theta} + 0.1e^{-i\theta} + \varepsilon & 0.02 + 0.1e^{i\theta} + 0.1e^{-i\theta} \\ 0.02 + 0.1e^{i\theta} + 0.1e^{-i\theta} & 1.02 + 0.1e^{i\theta} + 0.1e^{-i\theta} + \varepsilon \end{pmatrix} \\ \cdot \begin{pmatrix} -100e^{i\theta} - 20 \\ 100e^{i\theta} + 20 \end{pmatrix} \\ = 2(1 + \varepsilon)(100e^{i\theta} + 20)^2$$

and

$$b^\dagger Bb|_{z_1=0, z_2=0} = (-100e^{i\theta} - 20, 100e^{i\theta} + 20) \begin{pmatrix} 0 & 20 \\ 20 & 0 \end{pmatrix} \begin{pmatrix} -100e^{i\theta} - 20 \\ 100e^{i\theta} + 20 \end{pmatrix} \\ = -40(100e^{i\theta} + 20)^2.$$

Now $\varepsilon < 19$ implies that

$$(\bar{b}^\dagger Ab)(e^{i\theta}) < |(b^\dagger Bb)(e^{i\theta})| \quad \forall e^{i\theta} \in \mathbb{T},$$

so that (1) of Theorem 2.7 does not hold. We also have that

$$\omega(bBb^t) = 2 = 2\omega(a),$$

and (2) of Theorem 2.7 also fails. Then Theorems 2.7 and 2.2 imply that f^* is not a local directional optimizer.

Now we turn to proofs.

The following is a more general version of Theorem 2.2 in that the differentiability of $\Gamma(e^{i\theta}, z)$ at $z = f^*(e^{i\theta})$ is required only almost everywhere on \mathbb{T} . This weaker hypothesis and even weaker hypotheses are physically desirable for treating problems that arose in §5.

THEOREM 2.8. *Suppose that E is either H_N^∞ or A_N . Let $\Gamma : \mathbb{T} \times \mathbb{C}_N \rightarrow \mathbb{R}^+$ and $f^* \in E$ be given, such that*

- (1) *The function $g(\cdot) \triangleq \Gamma(\cdot, f^*(\cdot))$ is continuous on \mathbb{T} ;*
- (2) *$\Gamma(e^{i\theta}, z)$ is of class C^2 at $z = f^*(e^{i\theta})$ for almost all $e^{i\theta} \in \mathbb{T}$. Moreover, for such $e^{i\theta} \in \mathbb{T}$, the remainders $R_l(e^{i\theta}, z)$, $l = 1, 2$ of the order l Taylor expansion of $\Gamma(e^{i\theta}, f^*(e^{i\theta}) + z)$ about $z_0 = 0$ satisfy*

$$(2.4) \quad \sup_{\theta} |R_l(e^{i\theta}, z)/\|z\|_{\mathbb{C}_N}^l \rightarrow 0 \quad \text{as } z \rightarrow 0, l = 1, 2;$$

- (3) *The function $a(\cdot) = \partial\Gamma/\partial z(\cdot, f^*(\cdot))$ is bounded and bounded away from 0. If $E = A_N$, assume further that the set $\{e^{i\theta} \in \mathbb{T} : a \text{ is not continuous at } e^{i\theta}\}^-$ has linear Lebesgue measure 0;*

- (4) *There exists a continuous function $g_0 : \mathbb{T} \rightarrow \mathbb{C}^N$ such that*

$$\sup_{\theta} \left\| \frac{1}{\|a(\cdot)\|_{\mathbb{C}^N}} \cdot \bar{a}(\cdot) + g_0(\cdot) \right\|_{\mathbb{C}_N} < 1.$$

Then necessary conditions for f^ to be a directional optimizer for OPT_E are*

- (I) *g is constant;*
- (II) *$\chi^{-1} a$ has an analytic direction F ;*
- (III)

$$\sup_{\theta} \left\{ \overline{h(e^{i\theta})}^\dagger \frac{\partial^2 T}{\partial \bar{z} \partial z}(P(e^{i\theta})) h(e^{i\theta}) + \operatorname{Re} h(e^{i\theta})^\dagger \frac{\partial^2 \Gamma}{\partial z^2}(P(e^{i\theta})) h(e^{i\theta}) \right\} \geq 0$$

$$\forall h \in N(a) \cap E.$$

These conditions are also sufficient for f^ to be a strict directional optimizer if (III) is modified as to have strict inequality.*

Proof. Consider the function Γ_1 on $\mathbb{T} \times \mathbb{C}_N$ defined by

$$\begin{aligned} \Gamma_1(e^{i\theta}, z) &= g(e^{i\theta}) + 2 \operatorname{Re} \frac{\sqrt{g(e^{i\theta})}}{|a(e^{i\theta})|} a(e^{i\theta})^\dagger z + \|z\|_{\mathbb{C}_N}^2 \\ &= \left\| \frac{\sqrt{g(e^{i\theta})}}{\|a(e^{i\theta})\|_{\mathbb{C}^N}} \bar{a}(e^{i\theta}) + z \right\|_{\mathbb{C}_N}^2. \end{aligned}$$

If at least one of the hypotheses (I), (II) is not satisfied, then the function $0 \in A_N$ is not a solution to OPT_E for Γ_1 (Theorem A1 in Appendix A). Therefore, by Theorem A1 or Theorem A2, there exists a function $f' \in E$ such that

$$(2.5) \quad \sup_{\theta} \Gamma_1(e^{i\theta}, f'(e^{i\theta})) < \gamma^* \triangleq \sup_{\theta} \Gamma_1(e^{i\theta}, 0).$$

We claim that there exists an open set F that contains the set $P = \{e^{i\theta} \in \mathbb{T} : g(e^{i\theta}) = \gamma^*\}$ and a constant $\delta_0 \geq 0$, such that

$$(2.6) \quad \sup_{e^{i\theta} \in F} \operatorname{Re} a(e^{i\theta})^\dagger f'(e^{i\theta}) < -\delta_0.$$

To prove the claim, consider any decreasing sequence $\{F_n\}$ of open sets in \mathbb{T} such that $\cap F_n = P$. If the claim is false, then

$$(2.7) \quad m \left\{ e^{i\theta} \in F_n : \operatorname{Re} \frac{\sqrt{g(e^{i\theta})}}{|a(e^{i\theta})|} \cdot a(e^{i\theta})^\dagger f'(e^{i\theta}) > \frac{-1}{n} \right\} > 0 \quad \forall n \in \mathbb{N}.$$

Here mC is the Lebesgue measure of C . Choose $\delta > 0$ arbitrary and pick $N_\delta \in \mathbb{N}$ such that

$$(2.8) \quad g(e^{i\theta}) > \gamma^* - \delta \quad \forall e^{i\theta} \in F_n, n \geq N_\delta.$$

Then, from (2.7) and (2.8),

$$(2.9) \quad m \left\{ e^{i\theta} \in F_n : g + 2 \operatorname{Re} \frac{\sqrt{g(e^{i\theta})}}{|a(e^{i\theta})|} a(e^{i\theta})^\dagger f'(e^{i\theta}) + |f'(e^{i\theta})|^2 > \gamma^* - \delta - \frac{2}{n} \right\} > 0$$

$\forall n \geq N_\delta,$

i.e.,

$$(2.10) \quad \sup_{\theta} \Gamma_1(e^{i\theta}, f'(e^{i\theta})) \geq \gamma^* - \delta.$$

Since δ is arbitrary, it follows from (2.10) that

$$(2.11) \quad \sup_{\theta} \Gamma_1(e^{i\theta}, f'(e^{i\theta})) \geq \gamma^*,$$

which is impossible, by (2.5). This proves the claim.

Consider the following expansion, valid almost everywhere on \mathbb{T} :

$$(2.12) \quad \begin{aligned} \Gamma(e^{i\theta}, f^*(e^{i\theta}) + t(f'(e^{i\theta}) - f^*(e^{i\theta}))) &= \Gamma(e^{i\theta}, f^*(e^{i\theta})) \\ &+ 2t \operatorname{Re} a(e^{i\theta})^\dagger (h(e^{i\theta}) - f^*(e^{i\theta})) \\ &+ R_1(e^{i\theta}, t(f'(e^{i\theta}) - f^*(e^{i\theta}))). \end{aligned}$$

By (2.6) and hypothesis (2), there exist $t_0 > 0$ such that

$$(2.13) \quad \sup_{e^{i\theta} \in F} \Gamma(e^{i\theta}, f^*(e^{i\theta}) + t(f'(e^{i\theta}) - f^*(e^{i\theta}))) \leq \gamma^* - t\delta_0 \quad \forall t \in (0, t_0).$$

Also, since the open set F contains P and since $\Gamma(e^{i\theta}, f^*(e^{i\theta}))$ is continuous, there exists $t_1 \geq 0$ such that

$$(2.14) \quad \sup_{e^{i\theta} \in \mathbb{T}/F} \Gamma(e^{i\theta}, f^*(e^{i\theta}) + t(f'(e^{i\theta}) - f^*(e^{i\theta}))) < \gamma^* \quad \forall t \in (0, t_1).$$

Combining (2.13) and (2.14), we obtain that f^* is not a local directional optimizer, and this shows that (I) and (II) are necessary.

Let $h \in N(a)$. An order 2 expansion of $\Gamma(e^{i\theta} z)$ about $z = f^*(e^{i\theta})$ gives

$$\begin{aligned} 0 &\leq \sup_{\theta} \frac{1}{t^2} \{ \Gamma(e^{i\theta}, f^*(e^{i\theta}) + th(e^{i\theta})) - \gamma^* \} \\ (2.15) &\leq \sup_{\theta} 2 \left\{ \overline{h(e^{i\theta})}^{\dagger} \frac{\partial^2 \Gamma}{\partial \bar{z} \partial z}(e^{i\theta}, f^*(e^{i\theta})) h(e^{i\theta}) + \operatorname{Re} h(e^{i\theta})^{\dagger} \frac{\partial^2 \Gamma}{\partial z^2}(e^{i\theta}, f^*(e^{i\theta})) h(e^{i\theta}) \right\} \\ &\quad + \sup_{\theta} \frac{1}{t^2} R_2(e^{i\theta}, th(e^{i\theta})) \end{aligned}$$

Let $t \rightarrow 0$ in (2.15) and use (2.4) to obtain (III).

For sufficiency, assume (I)–(III) and that $f^* \in E$ is not a directional optimizer; i.e., there exists $h \in H_N^{\infty}$ and $t_0 > 0$ such that

$$\sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta}) + th(e^{i\theta})) < \gamma^* \quad \forall t \in (0, t_0).$$

We first analyze the case where $h \notin N(a)$.

LEMMA 2.9. *If $h \notin N(a)$, then $\sup_{\theta} \operatorname{Re} a(e^{i\theta})^{\dagger} h(e^{i\theta}) > 0$.*

Proof. If $\sup_{\theta} \operatorname{Re} a(e^{i\theta})^{\dagger} h(e^{i\theta}) \leq 0$, then, since (II) holds, for some $\lambda > 0$ and $F \in H_N^1$, $a = \lambda \chi F$. Therefore

$$(2.16) \quad \sup_{\theta} \operatorname{Re} \chi(e^{i\theta}) F(e^{i\theta})^{\dagger} h(e^{i\theta}) \leq 0,$$

and Corollary 4.8 in [G] implies that $\chi F^{\dagger} h$ is either an outer function or the function zero. Since $\chi F^{\dagger} h$ has nontrivial inner part, we must have that $F^{\dagger} h = 0$, i.e., $h \in N(a)$. \square

Set

$$(2.17) \quad s \triangleq \sup_{\theta} \operatorname{Re} a^{\dagger} h$$

and pick $t_0 > 0$ small enough so that the function R_1 in (2.12) satisfies

$$(2.18) \quad \sup_{\theta} \|R_1(e^{i\theta}, th(e^{i\theta}))\| < \frac{s}{2} t \quad \forall t \in (0, t_0).$$

Then, an order 1 expansion of $\Gamma(e^{i\theta}, z)$ about $z = f^*(e^{i\theta})$ and (2.18) imply that

$$(2.19) \quad \sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta}) + th(e^{i\theta})) \geq \gamma^* + st - \frac{s}{2} t > \gamma^* \quad \forall t \in (0, t_0).$$

Hence $h \in N(a)$ implies that h is not a descent direction. If how $h \in N(a)$, then strict inequality in (III) for the function h , an order 2 expansion of $\Gamma(e^{i\theta}, z)$ about $z = f^*(e^{i\theta})$, and hypothesis (2), imply that, for some $t_1 > 0$,

$$\sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta}) + th(e^{i\theta})) > \gamma^* \quad \forall t \in (0, t_1).$$

This proves sufficiency.

THEOREM 2.8'. *In addition to the hypotheses to Theorem 2.8, suppose that Γ is real symmetric, i.e., (2.1) holds, and that $f^* \in RH_N^{\infty}$. Modify condition (II) by requiring that $F \in RH_N^1$, and modify condition (III) so that $h \in N(\partial\Gamma/\partial z(\cdot, f^*(\cdot))) \cap RE$. Then (I)–(III) are necessary for f^* to be a local optimizer to OPT_{RE} .*

Proof of Theorem 2.6. For each $j = 1, \dots, N$, let P_j, Q_j be polynomials containing the zeros poles of a_j in \mathbb{D} , and the double zeros and double poles of P_j and Q_j on \mathbb{T} . Let P be the greatest common divisor of all the P_j 's and let Q be the least common multiple of the Q_j 's. Then we have that

$$(2.20) \quad a = \frac{P}{Q} b,$$

where b extends analytically to a neighborhood of \mathbb{D}^- and has no zeros in \mathbb{D}^- .

Now use the formula

$$(e^{i\theta} - z_0)(1 - \bar{z}_0 e^{i\theta}) = e^{i\theta} |e^{i\theta} - z_0|^2 \quad \text{if } z_0 \in \mathbb{D}$$

and

$$(e^{i\theta} - z_0)^2 = -e^{i\theta} |e^{i\theta} - z_0|^2 \quad \text{if } z_0 \in \mathbb{T}$$

to obtain a measurable function $\lambda : \mathbb{T} \rightarrow \mathbb{R}^+$ such that

$$(2.21) \quad \frac{P}{Q} = \lambda \chi^{w(P)-w(Q)} R,$$

where R is a rational function with no zeros on poles on the closed unit disk \mathbb{D}^- . Note that $w(P) = z(a)$, $w(Q) = p(a)$. Hence (2.20) and (2.21) imply that

$$(2.22) \quad w(P) - w(Q) \leq w(a).$$

Equality in (2.22) follows from representation (2.3) in Theorem 2.5.

To prove (2), let $h \in A_1$ be defined by $h = e^{-u^*+ui}$, where u is a C^1 -function such that

$$(2.23) \quad a_{j_0} = \chi^{\omega(a_{j_0})} e^{v+ui}$$

and u^* is the harmonic conjugate of u . Since $u \in C^1$, we have $h \in A_1$. Moreover, the extension of h to the closed unit disk has no zeros there. From (2.23) and (II), we have that

$$(2.24) \quad \lambda \chi^{\omega(a)} F_{j_0} = \chi^{\omega(a_{j_0})} h e^{v+u^*},$$

and, since $1/h \in A_1$, we conclude that $F_{j_0}/h \in H_1^1$. Now $\omega(a_{j_0}) < \omega(a)$ is impossible, since otherwise the function $\chi^{\omega(a)-\omega(a_{j_0})} F_{j_0}/h$ is in H_1^1 , but (2.23) says that this function is positive almost everywhere on \mathbb{T} .

Finally, if $\omega(a_{j_0}) = 1$, then $\omega(a) = 1$, and from (II) we have that $F_{j_0}^{-1} \in A_1$ and

$$\frac{a_j}{a_{j_0}} = \frac{\lambda \chi F_j}{\lambda \chi F_{j_0}} = F_j / F_{j_0},$$

i.e., $a_j/a_{j_0} \in A_1$. \square

The following result establishes smoothness of λ and F of (II) in Theorem 2.2, given a smooth $(\partial\Gamma/\partial z)(\cdot, f^*(\cdot))$.

PROPOSITION 2.10. *Assume the hypotheses of Theorem 2.2. If $\partial\Gamma/\partial z(\cdot, f^*(\cdot)) \in C^\infty$, then the function λ, F in (II) are, respectively, in C_1^∞ and C_N^∞ . Moreover, the function λ is never zero on \mathbb{T} if $\omega(a) = 1$.*

Proof. By the proof of Theorem 2.8, the function $h = 0$ is the best H_N^∞ -approximation to the continuous function $(1/\|a(\cdot)\|_{\mathbb{C}^N}) \cdot \bar{a}(\cdot)$; therefore $F_1 = \chi F$ is a singular vector for the Hankel operator of symbol $\bar{a}/\|a\|_{\mathbb{C}^N}$. By [HS], F is in $C_N^\infty(\mathbb{T})$. Moreover, the same proof shows that $a/\|a\| = F/\|F\|_{\mathbb{C}^N}$; i.e., the function $f_1\|F_1\|$ is in $C_N^\infty(\mathbb{T})$. Hence $\|F\|$ is also in C_1^∞ , and this implies that the zeros of F on \mathbb{T} , if any, have even order. Now each double zero of F on \mathbb{T} contributes 1 to $\omega(a)$, as we can see from

$$\frac{(e^{i\theta} - e^{i\theta_0})^2}{|e^{i\theta} - e^{i\theta_0}|^2} = -e^{i\theta}.$$

On the other hand, if $w(a) = 1$, then $w(F) = 0$, so that $\|F\|$ is in C_i^∞ . \square

Clearly, one requirement to simplify (III) is to know something about the structure of $N(\partial\Gamma/\partial z(\cdot, f^*(\cdot)))$. Our next lemma gives a parameterization of this set.

LEMMA 2.11. *Let $a \in C_N^1(\mathbb{T})$, $F \in H_N^1 \cap C_N^1(\mathbb{T})$ such that $a(e^{i\theta}) \neq 0$ for every θ and*

$$\frac{1}{\|a(e^{i\theta})\|_N} \cdot a(e^{i\theta}) = \frac{1}{\|F(e^{i\theta})\|} \cdot F(e^{i\theta}).$$

Then there exist functions $\{k_j\}_{j=1}^{N-1}$ in A_N such that

$$(2.25) \quad N(a) \cap A_N = A_1 k_1 + \cdots + A_1 k_{N-1}.$$

Also, for almost all $e^{i\theta} \in \mathbb{T}$, the set

$$\{k_1(e^{i\theta}), \dots, k_{N-1}(e^{i\theta})\}$$

spans in \mathbb{C}^N a linear subspace of (complex) dimension $N - 1$.

Proof. The first step is to find a matrix-valued function M in $H_{N \times L}^\infty$, $L \geq 1$, with C^1 entries whose boundary values $M(e^{i\theta})$ are rank $N - 1$ matrices and which satisfies $F^\dagger M = 0$. One such matrix M is

$$M = [M_1, M_2, \dots, M_N]$$

where

$$M_j = \begin{pmatrix} -F_1 & -F_2 & \cdots & -\hat{F}_j & \cdots & -F_N \\ F_j & 0 & \cdots & & & 0 \\ 0 & F_j & & \ddots & & \\ & 0 & & & & 0 \\ 0 & 0 & & 0 & & F_j \end{pmatrix}, \quad j = 1, \dots, N$$

and “ $\hat{}$ ” indicates that the corresponding column is suppressed. Clearly, the rank of M is at least $N - 1$, almost everywhere on \mathbb{T} since, for almost all $e^{i\theta} \in \mathbb{T}$, one of the F_j ’s is nonzero at $e^{i\theta}$. That the rank is not N follows from the fact that F is in the null space of M^\dagger . To remedy this, we take the inner-outer factorization $M^\dagger = \Theta Q$ of M with $\Theta \in H_{L \times N-1}^\infty$ and $Q \in H_{N-1 \times N}^\infty$ (cf. [RR1], [RR2], [Go]). Q is continuous because M is once differentiable. Now $\text{Range } Q(e^{i\theta}) = \text{Range } M(e^{i\theta})$, since $\Theta(e^{i\theta})$ is unitary, so $F^\dagger Q^\dagger = 0$, and this shows that $Q^\dagger A_{N-1} \subset N(a)$. Also, $\dim \text{Range } Q(e^{i\theta}) = N - 1$ for all θ , so, if $f \in N(a)$, then the function $h = (QQ^\dagger)^{-1}Qf$ is in A_{N-1} , and it satisfies $f = Q^\dagger h$. Therefore $N(a) \cap A_N = Q^\dagger A_{N-1}$. \square

Before proving Theorem 2.7, we restate (III) of Theorem 2.2.

LEMMA 2.12. Let Γ be of class C^2 and let $f^* \in A$ be a directional optimizer, such that an analytic direction F for χ^{-1} satisfies $F \in C^1$. Then (III) is equivalent to

$$(2.26) \quad \sup_{\theta} \{ \bar{f}^\dagger \tilde{A} f + \operatorname{Re} f^\dagger \tilde{B} f \} \geq 0 \quad \forall f \in A_{N-1}.$$

Here

$$\begin{aligned} \tilde{A} &\triangleq \left(\bar{k}_i^\dagger \frac{\partial^2 \Gamma}{\partial \bar{z} \partial z}(\cdot, f^*) k_j \right)_{i,j=1}^{N-1}, \\ \tilde{B} &\triangleq \left(k_i^\dagger \frac{\partial^2 \Gamma}{\partial z^2}(\cdot, f^*) k_j \right)_{i,j=1}^{N-1}, \end{aligned}$$

and $\{k_i\}_{i=1}^{N-1}$ are functions in A_N parameterizing $N(a) \cap A_N$ as in Lemma 2.11.

Proof. By Lemma 2.11, every $h \in N(a)$ has the form $h = \sum_{l=1}^{N-1} f_l k_l$ for some scalar valued functions $\{f_l\}_{l=1}^{N-1}$ in A_1 . Hence, for every θ , we have

$$\begin{aligned} & \overline{\left(\sum_{l=1}^{N-1} f_l(e^{i\theta}) k_l(e^{i\theta}) \right)}^\dagger A(e^{i\theta}) \left(\sum_{j=1}^{N-1} f_j(e^{i\theta}) k_j(e^{i\theta}) \right) \\ & + \operatorname{Re} \left\{ \left(\sum_{l=1}^{N-1} f_l(e^{i\theta}) k_l(e^{i\theta}) \right)^\dagger B(e^{i\theta}) \left(\sum_{j=1}^{N-1} f_j(e^{i\theta}) k_j(e^{i\theta}) \right)^\dagger \right\} \\ & = \sum_{l,j=1}^{N-1} \overline{f_l(e^{i\theta})} f_j(e^{i\theta}) \overline{k_l(e^{i\theta})}^\dagger A(e^{i\theta}) k_j(e^{i\theta}) + \operatorname{Re} \sum_{l,j=1}^{N-1} f_l(e^{i\theta}) f_j(e^{i\theta}) k_l(e^{i\theta})^\dagger B(e^{i\theta}) k_j(e^{i\theta}) \\ & = \overline{f(e^{i\theta})}^\dagger \tilde{A}(e^{i\theta}) f(e^{i\theta}) + \operatorname{Re} f(e^{i\theta})^\dagger \tilde{B}(e^{i\theta}) f(e^{i\theta}), \end{aligned}$$

and now the equivalence of (III) and (2.26) is clear. \square

Proof of Theorem 2.7. First, we claim that, if $F \triangleq (-f_2, f_1)^\dagger$, where $F_0 = (f_1, f_2)^\dagger$ is the analytic direction of $\chi^{-\omega(a)}a$, then

$$N(a) = \{\varphi F : \varphi \in H_1^\infty\}.$$

One inclusion is clear. The other inclusion we prove as follows: Since F is nonzero on \mathbb{D}^- by hypothesis, the Corona theorem asserts the existence of functions g_1, g_2 in H_1^∞ such that

$$(2.27) \quad f_1 g_1 + f_2 g_2 = 1,$$

where $F_0^\dagger = (f_1, f_2)$. Now pick $h \in N(a)$, so that

$$(2.28) \quad f_1 h_1 + f_2 h_2 = 0$$

and use (2.27) to obtain

$$h_1 = f_2 g_2 - g_1 f_2 h_2;$$

i.e., we have that, for $\varphi \triangleq g_1 h_2 - g_2 h_1$,

$$(2.29) \quad h_1 = -f_2 \varphi.$$

By substituting (2.29) into (2.28), we have $h_2 = f_1\varphi$. This proves the claim. Note that $b = \chi^{-\omega(a)}F$.

Now from the definition of b and from (2.1), we have

$$(2.30) \quad \begin{aligned} \tilde{A} &= \frac{1}{\lambda^2} \bar{b}^\dagger Ab, \\ \tilde{B} &= \frac{1}{\lambda^2 \chi^{2w(a)}} \cdot b^\dagger Bb \end{aligned}$$

and we can rewrite inequality (2.26) as

$$(2.31) \quad \sup_{\theta} \left\{ bAb^* |f|^2 + \operatorname{Re} \frac{f^2 bBb^t}{\chi^{2w(a)}} \right\} \geq 0.$$

Then the sufficiency of (i) and part (a) of (ii) of Theorem 2.7 is clear. Suppose that bBb^t is never zero and that $\omega(bBb^t)$ is odd, and pick $f \in A$ arbitrary. If f has a zero on \mathbb{T} , then (2.31) follows, while, if it does not have a zero, the winding number of $f^2 bBb^t / \chi^{2w(a)}$ about zero is a well-defined odd number. In this case, $\operatorname{Re} f^2 bBb^t / \chi^{2w(a)}$ is positive for some $e^{i\theta} \in \mathbb{T}$, and this shows that part (b) of (ii) of Theorem 2.7 is sufficient.

To prove necessity, assume that neither one of (i) and (ii) of Theorem 2.7 hold, i.e.,

$$(2.32) \quad b(e^{i\theta})A(e^{i\theta})b^*(e^{i\theta}) < |b(e^{i\theta})B(e^{i\theta})b^t(e^{i\theta})| \quad \forall e^{i\theta} \in \mathbb{T}$$

and that $w(bBb^t)$ is even and no greater than $2w(a)$. Then there exists $g_0 \in A_1$ with no zeros on \mathbb{T} so that

$$\frac{-\overline{(bBb^t)}}{|bBb^t|} \chi^{2w(a)} = \frac{g_0^2}{|g_0|^2}.$$

Hence

$$(2.33) \quad (bAb^*)|g_0|^2 + \operatorname{Re} \left\{ \frac{g_0^2(bBb^t)}{\chi^{2w(a)}} \right\} = (bAb^*)|g_0|^2 - |bBb^t| |g_0|^2,$$

and combining (2.32) and (2.33), we see that (III) fails. \square

2.2. A sufficient condition for true optima. The previous section treated local *directional* optima. Here we find conditions sufficient to guarantee that f^* is a local optimizer for OPT.

The question to be answered is whether (I)–(III) with strict inequality are enough to force f^* to be an optimizer (see Definition 1.1).

A strict local optimizer f^* has the property that sequences $\{f^k\}$ in A_N converging to f^* satisfy

$$(2.34) \quad \sup_{\theta} \Gamma(\cdot, f^k) > \sup_{\theta} \Gamma(\cdot, f^*) \quad \text{for } k \text{ large enough.}$$

This does not necessarily have to happen if f^* is merely a strict local directional optimizer, or if (almost equivalently) the hypotheses of Theorem 2.11 hold. In Theorem 2.13, we replace (III) by a stronger condition, which is sufficient to guarantee (2.34).

THEOREM 2.13. *Let Γ and f^* be as in the hypothesis of Theorem 2.2. For f^* to be a strict local optimizer, it is sufficient that, in addition to (I) and (II), the following condition holds:*

$$(2.35) \quad \text{For each } \theta \in [0, 2\pi) \text{ and every nonzero } z = (z_1, \dots, z_N) \in \mathbb{C}^N \text{ such that } a(e^{i\theta})^\dagger z = 0, \text{ the Hessian } H_\theta \text{ of } \Gamma(e^{i\theta}, \cdot) \text{ at } f^*(e^{i\theta}) \text{ satisfies } H_\theta[z, z] > 0.$$

Proof. Suppose that $f^* \in A_N$ satisfies (I), (II), and (2.35) and let $\{f^k\}$ be a sequence in $A_N \setminus \{0\}$ such that $h^k \triangleq f^k - f^* \rightarrow 0$ as $k \rightarrow \infty$; we see that (2.34) holds.

If $\{h^k\}$ is a subsequence such that $h^{k_j} \in N(a)$ for all j , then the Taylor expansion of $\Gamma(e^{i\theta}, \cdot)$ about $f^*(e^{i\theta})$ gives

$$(2.36) \quad \Gamma(e^{i\theta}, f^*(e^{i\theta}) + h^{k_j}(e^{i\theta})) = S^* + H_0[h^{k_j}(e^{i\theta}), h^{k_j}(e^{i\theta})] + E_2(e^{i\theta}, h^{k_j}(e^{i\theta})),$$

where $\sup_\theta |E_2(e^{i\theta}, h^{k_j}(e^{i\theta}))| \rightarrow 0$ as $j \rightarrow \infty$. Note that by (2.35), for each θ , there exists $\varepsilon(\theta) > 0$ such that $H_\theta[z, z] > \varepsilon(\theta)\|z\|_N^2$, if $a(e^{i\theta})z = 0$. If $\varepsilon \triangleq \inf_\theta \varepsilon(\theta)$, then $\varepsilon > 0$ by continuity of H_θ in θ . Hence the inequality

$$(2.37) \quad \Gamma(e^{i\theta}, f^*(e^{i\theta}) + h^k(e^{i\theta})) > s^* + \varepsilon\|h^k(e^{i\theta})\|_N^2 + E_2(e^{i\theta}, h^k(e^{i\theta}))$$

implies that $s^{k_j} > s^*$ for j large enough.

Now we can assume, without loss of generality, that, for every k , we have $h^k \notin N(a)$. By Lemma 2.14, which follows, for each θ there exists $\delta(e^{i\theta}) \in \mathbb{R}$ such that, for all $z \in \mathbb{C}^N$ with $a(e^{i\theta})z^\dagger \neq 0$,

$$(2.38) \quad H_\theta[z, z] > \delta(e^{i\theta})|a(e^{i\theta})^\dagger z|^2.$$

Note that negative $\delta(e^{i\theta})$'s are not excluded.

Pick δ_0 such that, for all θ , $\delta(e^{i\theta}) > \delta_0$. This is always possible if the numbers $\delta(e^{i\theta})$ are chosen in an optimal way, by continuity of H_θ in θ . Then, for each θ and all $z \neq 0$, $H_\theta[z, z] + |\delta_0||a(e^{i\theta})^\dagger z|^2$ is positive and, furthermore, is continuous in θ . Hence there exists $\varepsilon_0 > 0$ such that

$$(2.39) \quad H_\theta[z, z] + |\delta_0||a(e^{i\theta})^\dagger \cdot z|^2 > \varepsilon\|z\|^2 \quad \forall e^{i\theta} \in \mathbb{T}, z \in \mathbb{C}^N.$$

By Lemma 2.9, there exist elements $\theta_k, k \in \mathbb{N}$ such that

$$(2.40) \quad \operatorname{Re} a(e^{i\theta_k})^\dagger h^k(e^{i\theta_k}) = a(e^{i\theta_k})^\dagger h^k(e^{i\theta_k}) > 0.$$

Since $\|h^k\|_\infty \rightarrow 0$ as $k \rightarrow \infty$, there exists $k_1 \in \mathbb{N}$ such that

$$(2.41) \quad 2 > |\delta_0|a(e^{i\theta_k})^\dagger h^k(e^{i\theta_k}), \quad k \geq k_0.$$

Now consider for each k the Taylor expansion

$$(2.42) \quad \begin{aligned} \Gamma(e^{i\theta_k}, f^k(e^{i\theta_k})) &= \Gamma(e^{i\theta_k}, f^*(e^{i\theta_k})) + 2 \operatorname{Re} a(e^{i\theta_k})^\dagger h^k(e^{i\theta_k}) \\ &\quad + H_\theta[h^k(e^{i\theta_k}), h^k(e^{i\theta_k})] + E_2(e^{i\theta_k}, h^k(e^{i\theta_k})), \end{aligned}$$

where $\sup_\theta |E_2(e^{i\theta_k}, h^k(e^{i\theta_k}))|/\|h^k(e^{i\theta_k})\|_{\mathbb{C}^N}^2 \rightarrow 0$ as $k \rightarrow \infty$. Thus, from (2.39)–(2.41), we obtain

$$(2.43) \quad \begin{aligned} \Gamma(e^{i\theta_k}, f^k(e^{i\theta_k})) - s^* &\geq 2a(e^{i\theta_k})h^k(e^{i\theta_k}) + \varepsilon\|h^k(e^{i\theta_k})\|_N^2 \\ &\quad - |\delta_0| |a(e^{i\theta_k})h^k(e^{i\theta_k})|^2 + E_2(e^{i\theta_k}, h^k(e^{i\theta_k})) \\ &\geq \varepsilon\|h^k(e^{i\theta_k})\|_N^2 + E_2(e^{i\theta_k}, h^k(e^{i\theta_k})), \end{aligned}$$

and the conclusion of the theorem follows by taking k large enough in (2.43). \square

LEMMA 2.14. Let $b \in \mathbb{C}^N \setminus \{0\}$ and $Q[\cdot, \cdot]$ be a real-valued form on \mathbb{C}^1 that is real linear in each coordinate separately. If $Q[z, z] > 0$ for every $z \in \mathbb{C}^N A, z \neq 0$ such that $b^\dagger z = 0$, then there exists a real constant δ_0 such that

$$(2.44) \quad Q[z, z] > \delta_0 |b^\dagger z|^2 \quad \forall z \in \{z \in \mathbb{C}^N : b^\dagger z \neq 0\}.$$

The proof of Lemma 2.11 simply explains the fact that general real-valued quadratic form on \mathbb{C}^N satisfies an equality

$$Q[z, z] = \bar{z}^\dagger A z + \operatorname{Re}(z^\dagger B z) \quad \forall z \in \mathbb{C}^N,$$

where A is an $N \times N$ selfadjoint matrix and B is an $N \times N$ complex symmetric.

Proof of Lemma 2.14. Make a change of coordinates to assume without loss of generality that $b^\dagger = (b_1, 0, \dots, 0)$, so that, for any $z \in \mathbb{C}^N, z^\dagger \cdot b = 0$ if and only if $z^\dagger = (z_1, z_2, \dots, z_N)$ with $z_1 = 0$. By hypothesis, there exists $\delta > 0$ such that

$$(2.45) \quad Q[0, z_2, \dots, z_N], (0, z_2, \dots, z_N)] > 3\delta^2(|z_2|^2 + \dots + |z_N|^2)$$

with equality only if $z_i = 0$ for all i . From (2.45), we obtain that

$$(2.46) \quad Q[0, z_2, \dots, z_N], (0, z_2, \dots, z_N)] > 2 \sum_{i=2}^N \delta_i^2 |z_i|^2 + \operatorname{Re} \left(\sum_{i=2}^N \delta_i^2 z_i^2 \right).$$

If

$$(2.47) \quad Q[z, z] = \sum_{i=1}^N A_{ii} |z_i|^2 + \sum_{1 \leq i < j \leq N} \operatorname{Re} A_{ij} z_i \bar{z}_j + \sum_{1 \leq i \leq j \leq N} \operatorname{Re} B_{ij} z_i z_j,$$

where $a_{ii} > 0, A_{ij} = \overline{a_{ji}}$. Then by (2.46) we have

$$(2.48) \quad \begin{aligned} Q[z, z] &\geq A_{11} |z_1|^2 + \sum_{2 \leq j \leq N} \operatorname{Re} A_{1j} z_1 \bar{z}_j + \sum_{2 \leq j \leq N} \operatorname{Re} (B_{1j} z_1 z_j) \\ &\quad + 2 \sum_{j=2}^N \delta_j^2 |z_j|^2 + \operatorname{Re} \left(\sum_{j=2}^N \delta_j^2 z_j^2 \right) \\ &= A_{11} |z_1|^2 + \operatorname{Re} B_{11} z_1^2 + \sum_{j=2}^N S_j, \end{aligned}$$

where by definition

$$(2.49) \quad S_j \triangleq \operatorname{Re} A_{1j} z_1 \bar{z}_j + 2\delta_j^2 |z_j|^2 + \operatorname{Re} (B_{1j} z_1 z_j + \delta_j^2 z_j^2).$$

Hence, for each $j, 2 \leq j \leq N$, we have that

$$\begin{aligned} S_j &= \operatorname{Re} (A_{1j} - B_{1j}) z_1 \bar{z}_j + \delta_j^2 |z_j|^2 + \left| \frac{B_{1j} z_1}{2\delta} + \delta z_j \right|^2 \\ &\quad + \operatorname{Re} \left(\frac{B_{1j} z_1}{2\delta} + \delta z_j \right)^2 - \left| \frac{B_{1j}}{2\delta} \right|^2 |z_1|^2 - \operatorname{Re} \left(\frac{B_{1j}}{2\delta} \right)^2 z_1^2 \\ &= \left(- \left| \frac{B_{1j}}{2\delta} \right|^2 - \left| \frac{a_{1j} - B_{1j}}{2\delta} \right|^2 \right) |z_1|^2 - \operatorname{Re} \left(\frac{B_{1j}}{2\delta} \right)^2 z_1^2 + \left| \frac{A_{1j} - b_{1j}}{2\delta} z_1 + \delta z_j \right|^2 \\ &\quad + \left| \frac{B_{1j} z_1}{2\delta} + \delta z_j \right|^2 + \operatorname{Re} \left(\frac{B_{1j} z_1}{2\delta} + \delta z_j \right)^2, \end{aligned}$$

and we obtain the inequality

$$(2.50) \quad S_j \geq \left(-2 \left| \frac{B_{1j}}{2\delta} \right|^2 - \left| \frac{A_{1j} - B_{1j}}{2\delta} \right|^2 \right) |z_1|^2, \quad 2 \leq j \leq N.$$

Finally, (2.48) and (2.50) imply that

$$Q[z, z] \geq \left(a_{11} + |B_{11}| + \sum_{j=2}^N \left(-2 \left| \frac{B_{1j}}{2\delta} \right|^2 - \left| \frac{A_{1j} - B_{1j}}{2\delta} \right|^2 \right) \right) |z_1|^2. \quad \square$$

We close §2 with an example.

Example 2. Our objective is to make use of Theorem 2.7 and show that $f^* = (0, 0)$ is a strict local optimizer for the functions Γ_ε defined in Example 1, when $\varepsilon > 19$. It was shown in Example 1 that (I) and (II) hold for all $\varepsilon > 0$, and now let us fix θ . The Hessian of $\Gamma(e^{i\theta}, \cdot)$ at $f^*(e^{i\theta})$ and direction z is

$$(2.51) \quad \begin{aligned} H_\theta[z, z] = & 2(|e^{i\theta} + 0.1|^2 + 0.01 + \varepsilon)|z_1|^2 + 4(0.2 \cos \theta + 0.2) \operatorname{Re} z_1 \bar{z}_2 \\ & + 2(|e^{i\theta} + 0.1|^2 + 0.01 + \varepsilon)|z_2|^2 + 80 \operatorname{Re} z_1 z_2. \end{aligned}$$

The equation $a(e^{i\theta})^\dagger z = 0$ implies that, for some $z_1 \in \mathbb{C}$,

$$(2.52) \quad z = (z_1, -z_1).$$

Combine (2.51) and (2.52) to obtain

$$H_\theta[z, z] = 4(1 + \varepsilon)|z_1|^2 + 80 \operatorname{Re} z_1^2.$$

Since we assume $\varepsilon > 19$, it follows that $H_\theta[z, z] > 0$, for all $z \in \mathbb{C}^N$ such that $a(e^{i\theta})^\dagger z = 0$; i.e., (2.35) holds.

3. Nonuniqueness of solutions. One of the main open questions in the study of OPT is that of uniqueness of solutions for $N > 1$. In the case where $N = 1$, it is known.³

THEOREM 3.1 (see [HM]). *Let Γ be the class C^1 and let $\gamma^* > 0$ be a local optimum for OPT (3), such that the gradient $\partial\Gamma/\partial z(e^{i\theta}, z)$ never vanishes when $\Gamma(e^{i\theta}, z) = s^*$, the sets $S_\theta(s^*)$ are uniformly bounded in θ and, for all θ , diffeomorphic to the unit disk in C^1 . If a solution $f^* \in A_1$ to OPT for Γ exists, then it is unique (in A_1).*

For $N > 1$, it is known that there are functions Γ whose sets $S_\theta(s^*)$ are convex (but not strictly convex) where solutions to OPT are not unique. For example, $\Gamma(e^{i\theta}, z) = \max(|e^{i\theta} - z_1|, |z_2|)$ produces an OPT problem for which any

$$f_1(e^{i\theta}) = 0 \quad \text{and} \quad f_2 \in A_1, \quad |f_2(e^{i\theta})| \leq 1, \quad \forall e^{i\theta} \in \mathbb{T}$$

is a solution. This Γ is not smooth; however, smoothing it slightly still produces the same qualitative behavior in OPT.

THEOREM 3.2 (see [HH]). *If $N \geq 1$, Γ of class C^1 in z and both $\Gamma, \partial\Gamma/\partial z$ continuous in θ , and the sets $S_\theta(s^*)$ are strictly convex (uniformly in θ) and nondegenerate, then an H_N^∞ solution f^* to OPT for Γ exists and is unique.*

Since uniqueness is very desirable for computational success, it is worth considerable effort to find simple conditions on Γ that a priori guarantee uniqueness for OPT. One

³ The original version of Theorem 3.1 has stronger hypotheses. However, examination of its proof yields the formulation we give here.

appealing possibility that is consistent with all the evidence listed is that smooth Γ that are strictly plurisubharmonic⁴ in z yield a uniqueness for OPT. This is appealing because most Γ produced in engineering are at least approximable by such functions. Unfortunately, this pleasant prospect is not to be. This section is devoted to presenting a very well-behaved Γ when $N = 2$ for which OPT has two solutions.

Example 1. Our examples lie in a one parameter ($\varepsilon \geq 0$) family Γ_ε of objective functions defined in Example 1 of §2.

We prove the following proposition.

PROPOSITION 3.3. *The functions Γ_ε for $\varepsilon > 0$ are strictly plurisubharmonic for all $e^{i\theta} \in \mathbb{T}$ and*

(a) *For each ε , $0 < \varepsilon < 19$, the problem OPT has a strict local optimum s_ε^* at each of the constant functions*

$$f_\varepsilon^1 = (c_\varepsilon, -c_\varepsilon), \quad f_\varepsilon^2 = (-c_\varepsilon, c_\varepsilon),$$

where $c_\varepsilon = 5\sqrt{2(19 - \varepsilon)}$;

(b) *The functions $f_\varepsilon^1, f_\varepsilon^2$ belong to the same connected component of $S_\theta(s_\varepsilon^*)$ for all $\varepsilon < 19$ that are sufficiently close to 19.*

Proof. For all $\varepsilon > 0$ and each $(e^{i\theta}, z) \in \mathbb{T} \times \mathbb{C}^N$, the matrix

$$\begin{aligned} & \left(\left(\frac{\partial^2 \Gamma_\varepsilon(e^{i\theta}, z_1, z_2)}{\partial \bar{z}_i \partial z_j} \right) \frac{\partial^2 \Gamma}{\partial z} \right) \\ &= \begin{pmatrix} |e^{i\theta} + 0.1 + 0.1z_2|^2 + |0.1z_2 + 0.1|^2 + \varepsilon & (0.1\bar{z}_1 + 0.1)(e^{i\theta} + 0.1 + 0.1z_2) \\ (0.1\bar{z}_2 + 0.1)(e^{i\theta} + 0.1 + 0.1z_1) & + (e^{-i\theta} + 0.1\bar{z}_1 + 0.1)(0.1 + 0.1z_2) \\ + (e^{-i\theta} + 0.1\bar{z}^2 + 0.1)(0.1 + 0.1z_1) & |e^{i\theta} + 0.1 + 0.1z_2|^2 + |0.1z_2 + 0.1|^2 + \varepsilon \end{pmatrix} \end{aligned}$$

is strictly positive definite, which shows that $\Gamma(e^{i\theta}, \cdot)$ is strictly plurisubharmonic.

(a) We have that, for every θ ,

$$\begin{aligned} (3.1) \quad \Gamma_\varepsilon(e^{i\theta}, f_\varepsilon^1(e^{i\theta})) &= |100 - 0.1c_\varepsilon^2 + e^{i\theta}c_\varepsilon|^2 + |100 - 0.1c_\varepsilon^2 - e^{i\theta}c_\varepsilon|^2 + 2\varepsilon c_\varepsilon^2 \\ &= 2(100 - 0.1c_\varepsilon^2)^2 + 2c_\varepsilon^2(\varepsilon + 1). \end{aligned}$$

Hence (I) holds for f_ε^1 . We also have that, for all θ ,

$$\begin{aligned} (3.2) \quad \frac{\partial \Gamma_\varepsilon}{\partial z_1}(e^{i\theta}, f_\varepsilon^1(e^{i\theta})) &= (1 + \varepsilon)(5e^{i\theta} + 1), \\ \frac{\partial \Gamma_\varepsilon}{\partial z_2}(e^{i\theta}, f_\varepsilon^1(e^{i\theta})) &= (1 + \varepsilon)(5e^{i\theta} + 1). \end{aligned}$$

As it was done in Example 1, we find an analytic function $g \in A_1$ and a continuous $\lambda > 0$ such that

$$(3.3) \quad (1 + \varepsilon)(5e^{i\theta} + 1) = \lambda(e^{i\theta})\chi(e^{i\theta}) \cdot g(e^{i\theta}) \quad \forall e^{i\theta} \in \mathbb{T};$$

i.e., (II) holds for all $\varepsilon > 0$. To check (III), note first that, if $a_\varepsilon \triangleq (\partial \Gamma / \partial z)(\cdot, f_\varepsilon^1)$, then

$$A_2 \cap N(a_\varepsilon) = \{(h_1, h_2) \in A_2 : h_1 = -h_2\}$$

⁴ The function Γ is strictly plurisubharmonic in z means that

$$\bar{w}^\dagger \frac{\partial^2 \Gamma}{\partial \bar{z} \partial z}(e^{i\theta}, z)w > 0 \quad \forall w \in \mathbb{C}^N \setminus \{0\} \text{ such that } \frac{\partial \Gamma}{\partial z}(e^{i\theta}, z)^\dagger w = 0.$$

and that the Hessian of Γ_ε in z at f_ε^1 satisfies

$$\begin{aligned} H_\theta^\varepsilon[(z_1, z_2), (z_1, z_2)] &= 2(|e^{i\theta} + 0.1 - 0.1c_\varepsilon|^2 + |0.1 - 0.1c_\varepsilon|^2 + \varepsilon)|z_1|^2 \\ &\quad + 4 \operatorname{Re}((0.1c_\varepsilon + 0.1)(e^{i\theta} + 0.1 - 0.1c_\varepsilon) \\ &\quad + (e^{-i\theta} + 0.1 - 0.1c_\varepsilon))z_1\bar{z}_2 \\ &\quad + 2(|e^{i\theta} + 0.1 + 0.1c_\varepsilon|^2 + |0.1 + 0.1c_\varepsilon|^2 + \varepsilon)|z_2|^2 \\ &\quad + 4(\varepsilon + 1) \operatorname{Re} z_1 z_2. \end{aligned}$$

Hence, for $z_2 = -z_1$, we obtain, after some algebra,

$$H_\theta^\varepsilon[(z_1, -z_1), (z_1, -z_1)] = 4(19 - \varepsilon)|z_1|^2 + 4(\varepsilon + 1)(|z_1|^2 - \operatorname{Re} z_1^2),$$

so it is clear that (III) holds. Moreover, Theorem 2.2 applies, and OPT has a strict local solution at f_ε^1 for $\varepsilon < 19$. Note that, by symmetry, the same can be said of f_ε^2 .

(b) We have from (2.11) that

$$\left(\frac{\partial \Gamma_{\varepsilon_0}}{\partial z_1}(e^{i\theta}, 0), \frac{\partial \Gamma_{\varepsilon_0}}{\partial z_2}(e^{i\theta}, 0) \right) \neq 0 \quad \forall e^{i\theta} \in \mathbb{T}, \varepsilon \in [0, 19].$$

Then, by Lemma B1 of Appendix B, there exists in \mathbb{C}^N an open neighborhood V of 0 such that, for all θ and all $\varepsilon \in [0, 19]$, the sets

$$W(\theta, s) = \{z \in V : \Gamma_\varepsilon(e^{i\theta}, z) = s\}$$

are connected.

Now pick $\varepsilon < 19$ so close to 19 that both $f_\varepsilon^1(e^{i\theta})$ and $f_\varepsilon^2(e^{i\theta})$ lie in V for all θ . If we denote with s_ε the optimal value

$$s_\varepsilon = \sup_\theta \Gamma(e^{i\theta}, (c(\varepsilon), -c(\varepsilon))),$$

we have that

$$f_\varepsilon^i(e^{i\theta}) \in W(\theta, s_\varepsilon) \quad \forall e^{i\theta} \in \mathbb{T}, \quad i = 1, 2,$$

and this proves (b). \square

There are various applications of OPT in addition to the engineering ones that we emphasize in this paper. One application is in a branch of several complex variables; computing what is called a Kobayashi extremal for a particular domain \mathcal{D} in \mathbb{C}^n is a special case of OPT. (See [H2].) In particular, the problems produced by “strictly pseudoconvex” \mathcal{D} gives Γ that are strictly plurisubharmonic. Thus a Kobayashi extremal problem for such a \mathcal{D} with nonunique solution also is an example of the type given here. (L. Lempert (in a private communication) found such an example, but it is unpublished.)

4. Coordinate descent approaches to OPT. In standard \mathbb{R}^n optimization of nonlinear functions, it is a natural idea and fairly common in folklore to reduce the original problem to a sequence of one-dimensional problems through the technique of coordinate descent. As we see, the analogue of such a technique is seriously flawed in our setting. The explanation is based upon the difference in (II) of Theorem 2.2 between the cases where $N = 1$ and $N > 1$.

This section presents a simple (practical) test, Theorem 4.1, to determine when the natural coordinate descent algorithm for “solving” OPT stops. Fortunately, this stopping criterion compares directly to Theorem 2.2, which tells us when we are at a true optimum to OPT. As we see, the conditions of Theorem 4.1 are so much weaker than those of Theorem 4.1 that we propose the following conjecture.

CONJECTURE. *For generic Γ , coordinate descent does not obtain the true optimum. In other words, with probability one, coordinate descent gives the wrong answer.*

Now the coordinate descent CD algorithm is presented. Given $\Gamma : \partial\mathbb{D} \times \mathbb{C}^N \rightarrow \mathbb{R}^+$ and $f^0 = (A_1^0, f_2^0, \dots) \in A_N$, an update $f_1 \in A_N$ is obtained in the following way:

$$(CD1) \quad \text{Find } h_1 = h_1^* \in A_1 \text{ that minimizes } \sup_{\theta} \Gamma(\cdot, f_1^0 h_1, f_2^0, \dots, f_N^0);$$

$$(CD2) \quad \text{Find } h_2 = h_2^* \in A_2 \text{ that minimizes } \sup_{\theta} \Gamma(\cdot, f_1^0 + h_1^*, f_2^0 + h_2, f_3^0, \dots, f_N^0);$$

\vdots

$$(CDN) \quad \begin{array}{l} \text{Find } h_N = h_N^* \in A_N \text{ that minimizes} \\ \sup_{\theta} \Gamma(\cdot, f_1^0 + h_1^*, \dots, f_{N-1}^0 + h_{N-1}^*, f_N^0 + h_N). \end{array}$$

Then set $f^1 = f^0 + (h_1^*, \dots, h_N^*)$.

To analyze the CD algorithm, we write the mathematical statement of what it means for the algorithm to stop. We call $(f_1^*, f_2^*, \dots, f_N^*) = f^* \in A_N$ a *coordinate descent solution* to OPT, provided that

$$(CDS) \quad \begin{array}{l} \inf_{f_1} \sup_{\theta} \Gamma(e^{i\theta}, f_1, f_2^*, \dots, f_N^*) = \|\Gamma(\cdot, f^*\|_{\infty}, \\ \inf_{f_2} \sup_{\theta} \Gamma(e^{i\theta}, f_1^*, f_2, f_3^*, \dots, f_N^*) = \|\Gamma(\cdot, \dots, f^*)\|_{\infty}, \end{array}$$

and so forth.

THEOREM 4.1. *If $f^* \in A_N$ is such that $\partial\Gamma/\partial z_j(\cdot, f^*(\cdot))$ never equals zero on \mathbb{T} , for $j = 1, \dots, N$, then f^* is a CDS if and only if*

- (i) $\Gamma(\cdot, f^*(\cdot))$ is constant on \mathbb{T} ,
- (ii) (a) $\omega(\partial\Gamma/\partial z_j(\cdot, f^*(\cdot))) > 0, j = 1, \dots, N$, or, equivalently,
 (b) $\partial\Gamma/\partial z_j(\cdot, f^*(\cdot)) = \lambda_j(\cdot)\chi(\cdot)F_j(\cdot)$, where $\lambda_j : \mathbb{T} \rightarrow \mathbb{R}^+, j = 1, \dots, N$, are measurable functions and $F_j \in H_1^1$, for $j = 1, \dots, N$.

Proof. The proof follows from Theorem 2.1.

The key issue is the following: Does CD give solutions to OPT? To answer this, we compare this to conditions (I) and (II) of Theorem 2.1, which characterizes solutions to OPT. First, note that Theorem 4.1(i) and Theorem 2.2(I) are the same.

Now turn to condition (II). Here there is a big discrepancy. First, note that Theorem 4.1(ii) is exactly (II) of Theorem 2.2 with the (string) added condition that $\lambda_1 = \lambda_2 = \dots = \lambda_N$. Consequently, CD stops without even addressing the key optimality condition II in Theorem 2.2.

To get another perspective, suppose that the a_j have meromorphic continuations to a neighborhood of \mathbb{D} and consider the winding number conditions. Theorem 4.1 says that, inside \mathbb{D} ,

$$(\#\text{zeros} - \#\text{poles})(a_j) > 0, \quad j = 1, \dots, N.$$

Typically (see [M1]), we have

$$(4.1) \quad (\#\text{zeros} - \#\text{poles})(a_j) = 1, \quad j = 1, \dots, N,$$

while, for true optimum,

$$(4.2) \quad (\#\text{common zeros} - \text{total } \#\text{poles}(a_j)) > 0.$$

If we combine (4.2) with (4.3), we obtain that *all* zeros of the a_j are common (see §2).

Further evidence that condition (II) of Theorem 1 and Theorem 2.1 are worlds apart is supplied by simple computer experiments. The first case we tried is reported below.

Example 1. Let

$$\begin{aligned} \Gamma(e^{i\theta}, z_1, z_2) = & \left(\operatorname{re} \left(\frac{1}{e^{i\theta}} - z_1 \right) \right)^2 + 4 \left(\operatorname{im} \left(\frac{1}{e^{i\theta}} - z_1 \right) \right)^2 \\ & + \left(\operatorname{re} \left(\frac{1}{e^{i\theta}} - z_2 \right) \right)^2 + 0.25 \left(\operatorname{im} \left(\frac{1}{e^{i\theta}} - z_2 \right) \right)^2. \end{aligned}$$

Coordinate descent, initialized at $f^0(e^{i\theta}) = (0, 0)$, yields for this function optimal values $s_c^* \cong 2.54423$, attained at $f_c(e^{i\theta}) = (((-10 + \sqrt{73})/6)e^{i\theta}, 0)$. These values occur at step (CD1). On the other hand, Theorems 2.2 and 2.6 can be used to prove that f_c is not a local optimizer for OPT. In fact, $s^* \cong 2.09101$ is the (unique) optimal value, attained at

$$f^*(e^{i\theta}) \cong (-0.421325e^{i\theta}, 0.733838e^{i\theta}).$$

Example 2. Set

$$\begin{aligned} \Gamma(e^{i\theta}, z_z, z_2) = & (2 + 0.1 \cos \theta) \left(\operatorname{re} \left(\frac{1}{e^{i\theta} - 0.2} - z_1 \right) \right)^2 + 9 \left(\operatorname{im} \left(\frac{1}{e^{i\theta} - 0.2} - z_1 \right) \right)^2 \\ & + (4 - 0.2 \cos \theta) \left(\operatorname{re} \left(\frac{1}{e^{i\theta}} - z_2 \right) \right)^2 + \left(\operatorname{im} \left(\frac{1}{e^{i\theta}} - z_2 \right) \right)^2. \end{aligned}$$

The optimal value in OPT is $s^* \cong 5.28606$, while coordinate descent, initialized at $f^0 = (0, 0)$, completes (CD1) to finally get to a complete halt right at (CD2), giving $s_c^* \cong 7.14102$. Thus, in this example, there is approximately a 35% error in the answer.

5. The UNCOPT problem. In §2 we treated the problem OPT. As it was mentioned there, OPT is central to the problem of system design in the frequency domain. To formulate OPT in this context, complete knowledge of the part of the system that is given is required, to be able to determine what value the performance function Γ takes at specified values $e^{i\theta} \in \mathbb{T}$ and $z \in \mathbb{C}^N$. This is an idealized situation: typically, the engineer is confronted with the problem of design under the presence of unknown values for some of the parameters α that enter the model. However, he or she may know ranges through which α can vary. The standard jargon in control calls the parameter α the plant uncertainty.

In this situation, a performance function $\tilde{\Gamma}(e^{i\theta}, \alpha, z)$ is available, where α lives in a set $\mathcal{R}(e^{i\theta}) \subset \mathbb{R}^K$. Since the exact values of α are not known, a true worst-case analysis uses the function UNC in §1, $\Gamma(e^{i\theta}, z) \triangleq \sup_{\alpha \in \mathcal{R}(e^{i\theta})} \tilde{\Gamma}(e^{i\theta}, \alpha, z)$.

It is the purpose of this section to study this and other similar problems. More precisely, we state conditions for (local) optimality, when the function Γ is smooth, which involve in

their formulation only the function $\tilde{\Gamma}$ (and not Γ). For more details about the engineering application, we refer to [Doy], [H4], [HMer].

Let $\tilde{\Gamma}$ be a positive-valued function of $(e^{i\theta}, \alpha, z) \in \mathbb{T} \times \mathbb{R}^k \times \mathbb{C}_N$, such that all derivatives in α and z of order up to 3 exist everywhere and are continuous in $e^{i\theta}$. Let $g : \mathbb{T} \times \mathbb{R}^K \rightarrow \mathbb{R}$ be a C^2 -function, such that the sets

$$(5.1) \quad \mathcal{R}(e^{i\theta}) \triangleq \{\alpha \in \mathbb{R}^k / g(e^{i\theta}, \alpha) \leq 0\}$$

have nonempty interior and satisfy

$$(5.2) \quad \alpha \in \partial \mathcal{R}(e^{i\theta}) \Rightarrow \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha) \neq 0.$$

By UNCOPT, we denote the problem of finding solutions to either one of the statements UNCOPT-I and UNCOPT-S':

$$(UNCOPT-S') \quad \inf_{f \in A_N} \sup_{\theta} \sup_{\substack{\alpha \in \mathbb{R}^k \\ g(e^{i\theta}, \alpha) \leq 0}} \tilde{\Gamma}(e^{i\theta}, \alpha, f(e^{i\theta})),$$

$$(UNCOPT-I) \quad \inf_{f \in A_N} \sup_{\theta} \inf_{\substack{\alpha \in \mathbb{R}^k \\ g(e^{i\theta}, \alpha) \leq 0}} \tilde{\Gamma}(e^{i\theta}, \alpha, f(e^{i\theta})).$$

These problems split naturally and cleanly into two separate problems:

$$(5.3a) \quad \Gamma(e^{i\theta}, z) = \sup_{\alpha \in \mathcal{R}(e^{i\theta})} \tilde{\Gamma}(e^{i\theta}, \alpha, z)$$

for UNCOPT-S', or

$$(5.3b) \quad \Gamma(e^{i\theta}, z) = \inf_{\alpha \in \mathcal{R}(e^{i\theta})} \tilde{\Gamma}(e^{i\theta}, \alpha, z)$$

for UNCOPT-I. Then UNCOPT-S' and UNCOPT both become OPT for Γ . Indeed, it is convenient to define what we mean by solution to UNCOPT and UNCOPT-I directly in terms of OPT. We now formalize this.

DEFINITION 5.1. Let $\alpha^* : \mathbb{T} \rightarrow \mathbb{R}^k$ measurable and $f^* \in A_N$ and suppose that $\alpha^*(e^{i\theta}) \in \mathcal{R}(e^{i\theta})$, for all $e^{i\theta} \in \mathbb{T}$. Then we call (f^*, α^*) a feasible pair.

DEFINITION 5.2. Let (f^*, α^*) be a feasible pair. Then (f^*, α^*) is a local solution to UNCOPT-S' (respectively, UNCOPT-I), if

(1) f^* is a strict (local) directional minimizer to OPT for the function Γ in (5.3a) (respectively, (5.3b));

(2) for almost all $e^{i\theta} \in \mathbb{T}$, $\alpha^*(e^{i\theta})$ is an optimizer for $\tilde{\Gamma}(e^{i\theta}, \alpha, f^*(e^{i\theta}))$; i.e.,

$$(5.4) \quad \Gamma(e^{i\theta}, f^*(e^{i\theta})) = \tilde{\Gamma}(e^{i\theta}, \alpha^*(e^{i\theta}), f^*(e^{i\theta})).$$

Thus we obtain essentially by definition our first and most basic recipe for characterizing solutions.

RECIPE. To check if the feasible pair (f^*, α^*) solves UNCOPT locally,

(1) use the classical Kuhn-Tucker conditions (see [GMW]) to check if α^* optimizes $\Gamma(e^{i\theta}, \cdot, f^*(e^{i\theta}))$;

(2) use §2 to check that f^* is a solution to OPT for Γ given by (5.3). This is one of the main motivations for §2.

One important shortcoming of the recipe is that an explicit formula for Γ is not normally available. Therefore the results of §2 cannot be applied directly. We would like a test of optimality expressed directly in terms of $\tilde{\Gamma}$ rather than the secondary function Γ . Under what basically are smoothness assumptions for Γ , this can be carried out. That is the main objective of this section.

The smoothness hypothesis on Γ is not a trivial one. It is easy to find examples where this is violated (see comments following Theorem 5.7).

We caution the reader that local optima are not adequate to ensure robustness in engineering designs. Consequently, we must compute enough local optima to be sure of finding a global one.

Our goal is to characterize optima α^*, f^* directly in terms of $\tilde{\Gamma}$. A most naive approach would be to set

$$\Gamma_1(e^{i\theta}, z) = \tilde{\Gamma}(e^{i\theta}, \alpha^*(e^{i\theta}), z),$$

assume the same function f^* as solution, and then proceed. Unfortunately, the assumption is false, the conclusion it gives for condition III is wrong, and so the issue is much more difficult. We must consider that a solution α^* to (5.3) depends on both θ and z . Indeed, it is by applying this that we get our main results, below. First, we provide some notation.

Let (f^*, α^*) be an admissible pair. The following statements (I') and (II') are essential to the characterization of solutions to UNCOPT:

(I') $\tilde{\Gamma}(\cdot, \alpha^*(\cdot), f^*(\cdot))$ is constant on \mathbb{T} ;

(II') $\chi^{-1}(\cdot) \partial \tilde{\Gamma} / \partial z(\cdot, \alpha^*(\cdot), f^*(\cdot))$ has an analytic direction.

Statements (I') and (II') duplicate in an obvious way statements (I) and (II) of §2. There is also, under appropriate hypotheses, another statement involving second-order derivatives of $\tilde{\Gamma}$ (see III' in Theorem 5.7), which is more difficult to check than (I') and (II'). This we introduce later.

Note that (II') makes perfect sense for cases in which $\partial \tilde{\Gamma} / \partial z(\cdot, \alpha^*(\cdot), f^*(\cdot))$ is not continuous. In fact, we believe that (I') and (II') are necessary for optimality in many cases that are not covered by the theorems proved in this section. We analyze this in a later paper. We certainly recommend in practical situations considering both (I') and (II') as necessary, even if the hypotheses of theorems do not guarantee necessity for optimality.

DEFINITION 5.3. A feasible pair (f^*, α^*) is an *interior pair* (respectively, *boundary pair*), provided that there exists $\delta > 0$ such that $g(e^{i\theta}, \alpha^*(e^{i\theta}), f^*(e^{i\theta})) \leq \delta < 0$ almost everywhere on \mathbb{T} (respectively, $g(e^{i\theta}, \alpha^*(e^{i\theta}), f^*(e^{i\theta})) = 0$ almost everywhere on \mathbb{T}).

DEFINITION 5.4. An interior pair (f^*, α^*) is called *regular* if $\partial^2 g / \partial \alpha^2(e^{i\theta}, \alpha^*(e^{i\theta}))$ is invertible almost everywhere on \mathbb{T} , with inverse essentially bounded in norm on \mathbb{T} . A boundary pair is *regular* if the Hessian $(\partial^2 g / \partial u^2)(e^{i\theta}, \alpha^*(e^{i\theta}))$ when compressed to the tangent plane to $\partial \mathcal{R}(e^{i\theta})$ at $\alpha^*(e^{i\theta})$ is invertible (with inverse bounded uniformly in θ).

In other words, $D(e^{i\theta})^\dagger \partial^2 g / \partial \alpha^2(e^{i\theta}, \alpha^*(e^{i\theta})) D(e^{i\theta})$ is invertible almost everywhere on \mathbb{T} , with inverse essentially bounded in norm on \mathbb{T} . Here D is any $k \times k-1$ matrix-valued measurable function, such that $D(e^{i\theta})$ has as column vectors that form an orthonormal basis for the space $\{\beta \in \mathbb{R}^k : (\partial \tilde{\Gamma} / \partial \alpha)(P(e^{i\theta}))^\dagger \cdot \beta = 0\}$.

THEOREM 5.5. Let (f^*, α^*) be a regular interior or boundary pair such that α^* is continuous on \mathbb{T} and $(\partial \tilde{\Gamma} / \partial z)(\cdot, \alpha^*(\cdot), f^*(\cdot))$ never equals zero on \mathbb{T} . If (f^*, α^*) is a solution to UNCOPT, then (I) and (II) hold.

Conversely, if (I) and (II) hold for the regular interior pair (f^*, α^*) and if $\tilde{\Gamma}(e^{i\theta}, \alpha^*(e^{i\theta}), \cdot)$ is locally strictly convex at $z = f^*(e^{i\theta_0})$ for some $e^{i\theta_0} \in \mathbb{T}$, then (f^*, α^*) solves UNCOPT-S'.

One basic observation is that a given boundary pair that solves UNCOPT is also a solution to the problem obtained by constraining α to satisfy $g(e^{i\theta}, \alpha) = 0$. Since the analysis is always local, many results that apply to solutions that are interior pairs, after reparametrization, could be stated for the boundary case. This is the case in the next theorem. If (f^*, α^*) is a fixed feasible pair, we use the notation $P(e^{i\theta}) = (e^{i\theta}, \alpha^*(e^{i\theta}), f^*(e^{i\theta}))$. Also, by $\partial^2 \tilde{\Gamma} / \partial z \partial \alpha (P(e^{i\theta}))$, we denote the $N \times K$ matrix, with (l, j) entry $(\partial^2 \tilde{\Gamma} / \partial z_l \partial \alpha_j)(P(e^{i\theta}))$.

THEOREM 5.6. *Suppose that Γ, g , and (f^*, α^*) satisfy the hypotheses of Theorem 5.5 with the possible exception of convexity. Assume further that (f^*, α^*) is an interior pair. Then necessary conditions for (f^*, α^*) to solve UNCOPT are (I') and (II') as in Theorem 5.5, and for every $h \in N(\partial \tilde{\Gamma} / \partial z(\cdot, \alpha^*(\cdot), f^*(\cdot))) \cap A_N$,*

$$\begin{aligned} \sup_{\theta} \left\{ \overline{h(e^{i\theta})}^\dagger \left(\frac{-\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha} (P(e^{i\theta})) \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2} (P(e^{i\theta}))^{-1} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha} (P(e^{i\theta}))^\dagger + \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial z} (P(e^{i\theta})) \right) h(e^{i\theta}) \right. \\ \left. + \operatorname{Re} h(e^{i\theta})^\dagger \left(\frac{-\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha} (P(e^{i\theta})) \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2} (P(e^{i\theta}))^{-1} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha} (P(e^{i\theta}))^\dagger \right. \right. \\ \left. \left. + \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial z} (P(e^{i\theta})) \right) h(e^{i\theta}) \right\} \\ \geq 0. \end{aligned} \quad (5.5)$$

If strict inequality is placed in (5.5), then these conditions are also sufficient.

Suppose that (f^*, α^*) is a boundary pair with α^* continuous on \mathbb{T} and that $\tilde{\Gamma}$ has been reparameterized as

$$\tilde{\Gamma}(e^{i\theta}, \beta, z) = \tilde{\Gamma}(e^{i\theta}, w(e^{i\theta}, \beta), z),$$

where

$$w(e^{i\theta}, \cdot) : \mathbb{R}^{k-1} \rightarrow \partial \mathcal{R}(e^{i\theta}) \subset \mathbb{R}^k$$

is a smooth (on $\mathbb{T} \times \mathbb{R}^{k-1}$) parameterization of $\partial \mathcal{R}(e^{i\theta})$ valid near $\alpha^*(e^{i\theta})$. Then replace $\tilde{\Gamma}$ by $\tilde{\Gamma}$ and $\partial / \partial \alpha$ by $\partial / \partial \beta$ in (5.5) to obtain (5.5'). Thus the interior case theorem, above, holds for the boundary case, provided that (5.5') replaces (5.5).

Theorem 5.6 in the case where $N = 2$ can be reduced to a practical test, in identical manner to what was done in §2, Theorem 2.7.

We now present what is the most general result of this section. Assumptions require a fairly detailed knowledge of the function Γ arising from $\tilde{\Gamma}$ as in (5.3). This is hypothesis (iv) in the following theorem. Recall that, if v is in \mathbb{C}^N , then $\operatorname{Re} v$ is the column vector with entries $\operatorname{Re} v_l, l = 1, \dots, N$. Also, if α is an \mathbb{R}^k -valued function on \mathbb{C}^N , then, by $\partial \alpha / \partial z(\cdot)$, we denote a $K \times N$ matrix with (j, l) entry given by $\partial \alpha_j / \partial z_l(\cdot)$.

THEOREM 5.7. *Let (f^*, α^*) be a pair of admissible functions, such that $\partial \tilde{\Gamma} / \partial z(\cdot, \alpha^*(\cdot), f^*(\cdot))$ never equals zero on \mathbb{T} . Suppose that there exists an open set $V \subset \mathbb{C}^N$ and a function $\alpha : \mathbb{T} \times V \rightarrow \mathbb{R}^k$ such that*

- (i) $g(e^{i\theta}, \alpha(e^{i\theta}, z)) \leq 0, \forall z \in V, e^{i\theta} \in \mathbb{T}$;
- (ii) *For each $e^{i\theta} \in \mathbb{T}$, the function $\alpha(e^{i\theta}, \cdot)$ is C^2 on V . Also, $\alpha(e^{i\theta}, \cdot)$ and its derivatives in z are continuous in $e^{i\theta} \in \mathbb{T}$;*
- (iii) $\alpha(e^{i\theta}, 0) = \alpha^*(e^{i\theta}), \forall e^{i\theta} \in \mathbb{T}$;
- (iv) $\tilde{\Gamma}(e^{i\theta}, \alpha(e^{i\theta}, z), f^*(e^{i\theta}) + z) = \Gamma(e^{i\theta}, f^*(e^{i\theta}) + z), \forall e^{i\theta} \in \mathbb{T}, z \in V$.

For each $e^{i\theta} \in \mathbb{T}$, let $\eta = \eta(e^{i\theta}, 0)$ be the unique solution to

$$\frac{\partial \tilde{\Gamma}}{\partial \alpha} (P(e^{i\theta})) - \eta \frac{\partial g}{\partial \alpha} (e^{i\theta}, \alpha^*(e^{i\theta})) = 0.$$

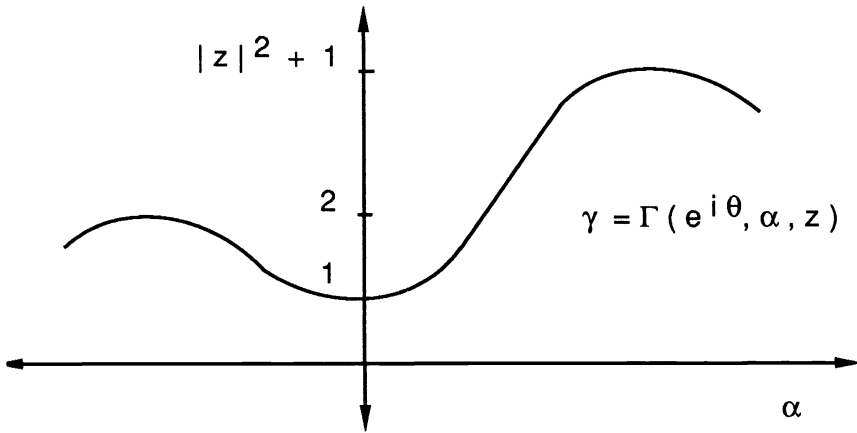


FIG. 5.1

Typically, η is referred to as a Lagrange multiplier. Then conditions (I'), (II'), and condition (III'), below, are necessary for (f^*, α^*) to be a solution to UNCOPT.

(III') For every $h \in N((\partial \tilde{\Gamma} / \partial z)(\cdot, \alpha^*(\cdot), f^*(\cdot)))$,

$$\sup_{\theta} \left\{ 4 \operatorname{Re} \left(h(e^{i\theta})^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \right) 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger h(e^{i\theta}) \right. \\ + \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger h(e^{i\theta}) \right)^\dagger \left(\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta}, \cdot)) \right) \\ \cdot \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger h(e^{i\theta}) \right) \\ \left. + 2 \overline{h(e^{i\theta})} \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial z}(P(e^{i\theta})) h(e^{i\theta}) + 2 \operatorname{Re} h(e^{i\theta})^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z^2}(P(e^{i\theta})) h(e^{i\theta}) \right\} \geq 0.$$

If strict inequality is considered in (III'), then the three conditions above are also sufficient.

A fragile hypothesis in this theorem is the assumption that $\alpha(\cdot)$ is a differentiable function. While it often might hold for interior or boundary pairs (f^*, α^*) , it seems unlikely to survive a case where α moves from the interior to the boundary. Even interior pairs (f^*, α^*) may occur with discontinuous α^* . We illustrate this with a pictorial example, in which the function Γ arising from (5.3) is continuous but not differentiable.

The plot of a function $\tilde{\Gamma}(e^{i\theta}, \alpha, z)$ with z and $e^{i\theta}$ fixed is shown in Figure 5.1. We assume that for this function $\tilde{\Gamma}$ the picture in the left half plane in Figure 5.1 does not change with z , while the peak in the right half plane has height given by $|z|^{2+1}$.

Hence, for this function $\tilde{\Gamma}$, we have that

$$\Gamma(e^{i\theta}, z) = \sup_{\alpha} \tilde{\Gamma}(e^{i\theta}, \alpha, z) = \begin{cases} 2 & \text{if } |z| \leq 1, \\ |z|^2 + 1 & \text{if } |z| > 1, \end{cases}$$

and it is clear that Γ is not differentiable in z , if $|z| = 1$.

A practical inconvenience with the previous result is that it requires us to know not only $\alpha(e^{i\theta}, f^*(e^{i\theta}))$, but also $\partial\alpha/\partial z(e^{i\theta}, 0)$. Fortunately, Theorem 5.6 has no dependence on $\partial\alpha/\partial z(e^{i\theta}, 0)$, so this is one advantage of it.

Basic for proving Theorem 5.7 is the following expansion:

$$\begin{aligned} \tilde{\Gamma}(e^{i\theta}, \alpha, f^*(e^{i\theta}) + tz) &= \tilde{\Gamma}(P(e^{i\theta})) + 2 \operatorname{Re} \frac{\partial \tilde{\Gamma}}{\partial z}(P(e^{i\theta}))^\dagger(tz) \\ &\quad + \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger(\alpha - \alpha^*(e^{i\theta})) \\ &\quad + \frac{1}{2} \left\{ 2 \operatorname{Re} tz^t \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha \partial z}(P(e^{i\theta}))tz \right. \\ &\quad + (\alpha - \alpha^*(e^{i\theta}))^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta}))(\alpha - \alpha^*(e^{i\theta})) \\ &\quad + 2t^2 \bar{z}^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial z}(P(e^{i\theta}))z + 2t^2 \operatorname{Re} \frac{\partial^2 \tilde{\Gamma}}{\partial z^2}(P(e^{i\theta}))z \Big\} \\ &\quad + E_2(e^{i\theta}, \alpha, tz) \quad \forall e^{i\theta} \in \mathbb{T}, z \in V, t \in (0, 1), \end{aligned} \tag{5.6}$$

where the function E_2 is continuous on $\mathbb{T} \times V$ and

$$\sup_{\theta} \frac{1}{t^2}, E_2(e^{i\theta}, \alpha(e^{i\theta}, tz), tz) \rightarrow 0 \quad \text{as } t \rightarrow 0. \tag{5.7}$$

LEMMA 5.8. *Under the hypotheses of Theorem 5.7,*

$$\frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger \cdot \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger = 0 \quad \forall (e^{i\theta}) \in \mathbb{T}.$$

Proof. Fix $e^{i\theta} \in \mathbb{T}$. Then hypothesis (iv) implies that

$$\frac{d}{dt} \tilde{\Gamma}(e^{i\theta}, \alpha(e^{i\theta}, tw), f^*(e^{i\theta}))|_{t=0} = 0 \quad \forall w \in V,$$

i.e.,

$$\begin{aligned} \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger \cdot 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger w \\ = 2 \operatorname{Re} \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger w = 0 \quad \forall w \in \mathbb{C}^N, e^{i\theta} \in \mathbb{T}. \end{aligned} \tag{5.8}$$

Set $w = \overline{(\partial\alpha/\partial z)(e^{i\theta}, 0)(\partial\tilde{\Gamma}/\partial\alpha)(P(e^{i\theta}))}$ in (5.8) to obtain

$$\left\| \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0) \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta})) \right\|_{\mathbb{C}_N} = 0 \quad \forall e^{i\theta} \in \mathbb{T}. \quad \square$$

LEMMA 5.9. *Assume the hypotheses of Theorem 5.7 and let $(f^* \alpha^*)$ be a solution to UNCOPT. Then $h^* = 0$ is a directional minimizer for $\sup_{\theta} \Gamma_1(e^{i\theta}, h(e^{i\theta}))$ over A_N , where Γ_1 is defined as*

$$\Gamma_1(e^{i\theta}, z) \triangleq \tilde{\Gamma}(P(e^{i\theta})) + 2 \operatorname{Re} \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger z + \|z\|_{\mathbb{C}^N}^2. \tag{5.9}$$

Proof. Combine expansion (5.6) with Lemma 5.8 to obtain a function $G_1 : \mathbb{T} \times \mathbb{R}^k \times \mathbb{C}^N \rightarrow \mathbb{R}$ such that

$$(5.10) \quad \frac{1}{t} \sup_{\theta} |G_1(e^{i\theta}, \alpha(e^{i\theta}, tz), tz)| \rightarrow 0 \quad \text{as } t \rightarrow 0 \quad \forall z \in \mathbb{C}^N$$

and

$$(5.11) \quad \begin{aligned} & \tilde{\Gamma}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz) \\ &= \tilde{\Gamma}(P(e^{i\theta})) + 2t \operatorname{Re} \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger z + G_1(e^{i\theta}, \alpha(e^{i\theta}, tz), tz). \end{aligned}$$

The remainder of the proof is similar to the first part of the proof of Theorem 2.8. \square

The necessity of (I') and (II') follows immediately from the proof of Theorem 2.8 and from Lemma 5.9. Now we move to (III'). We need the following lemma.

LEMMA 5.10. *Under the hypotheses of Theorem 5.7, if the pair (f^*, α^*) is a solution to UNCOPT, then there exists $t_1 > 0$ and a continuous function $h_0 : \mathbb{T} \times t_1 V \rightarrow \mathbb{R}$ such that*

$$(5.12) \quad \begin{aligned} & \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger (\alpha(e^{i\theta}, tz) - \alpha^*(e^{i\theta})) \\ &= -\eta(e^{i\theta}, z) \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right)^\dagger \frac{\partial^2 g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \\ & \quad \cdot \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right) t^2 \\ & \quad + R_0(e^{i\theta}, tz) \quad \forall t \in (0, t_1), z \in V, e^{i\theta} \in \mathbb{T} \end{aligned}$$

and

$$\sup_{\theta} \frac{R_0(e^{i\theta}, tz)}{t \|z\|_{\mathbb{C}^N}} \rightarrow 0 \quad \text{as } t \rightarrow 0.$$

Proof. If (f^*, α^*) is a regular interior point, then $(\partial \tilde{\Gamma} / \partial \alpha)(P(e^{i\theta})) = 0$ for all $e^{i\theta} \in \mathbb{T}$. Set $R_2 = 0$ and $t = 1$ to obtain the conclusion of the theorem.

Now assume that (f^*, α^*) is a regular boundary point. We have, from the expansion of $\alpha(e^{i\theta}, tz)$ about $t = 0$, the definition of $\eta(\cdot, 0)$, and from Lemma 5.8 that, for some $t' > 0$,

$$(5.13) \quad \begin{aligned} & \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger (\alpha(e^{i\theta}, tz) - \alpha^*(e^{i\theta})) \\ &= -\eta(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha^*(e^{i\theta}))^\dagger \left(\frac{d}{dt^2} \alpha(e^{i\theta}, tz) \Big|_{t=0} \right) \frac{t^2}{2} + R_2(e^{i\theta}, tz) \\ & \quad \forall z \in V, \quad t \in (0, t'), \quad e^{i\theta} \in \mathbb{T}, \end{aligned}$$

where R_2 is a continuous function satisfying $\sup_{\theta} (1/t \|z\|_{\mathbb{C}^N}) \cdot R_2(e^{i\theta}, tz) \rightarrow 0$ as $t \rightarrow 0$. Since for all $e^{i\theta} \in \mathbb{T}$, $\partial \tilde{\Gamma} / \partial w(P(e^{i\theta})) \neq 0$, then there exists $t_1 > 0$ such that

$$(5.14) \quad g(e^{i\theta}, \alpha(e^{i\theta}, tz)) = 0 \quad \forall t \in (0, t_1), \quad z \in V.$$

Note that, by compactness of \mathbb{T} , t_1 can be chosen to be independent of $e^{i\theta}$. Differentiating in (5.14) with respect to t , we have

$$(5.15) \quad \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha(e^{i\theta}, tz))^\dagger 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, tz)^\dagger z = 0 \quad \forall t \in (0, t_1), \quad z \in V.$$

Differentiating again, we obtain from (5.15)

$$(5.16) \quad \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right)^\dagger \frac{\partial^2 g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \\ + \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha^*(e^{i\theta}))^\dagger \left(\frac{d}{dt^2} \alpha(e^{i\theta}, tz) \Big|_{t_0=0} \right) = 0.$$

We combine (5.16) and (5.13) and set $R_0 = R_2$ to finish the proof of the lemma. \square
Combine (5.8) with Lemma 5.10 to obtain the expansion

$$(5.17) \quad \tilde{\Gamma}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz) \\ = \tilde{\Gamma}(P(e^{i\theta})) + 2t \operatorname{Re} \frac{\partial \tilde{\Gamma}}{\partial \alpha}(P(e^{i\theta}))^\dagger z \\ + \frac{t^2}{2} \left\{ 4 \operatorname{Re} z^t \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right. \\ + \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right)^\dagger \left(\frac{\partial \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \right) \\ \cdot \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right) \\ \left. + \bar{z}^\dagger \frac{\partial^2 \Gamma}{\partial \bar{z} \partial z}(P(e^{i\theta})) z + 2 \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z^2}(P(e^{i\theta})) z \right\} \\ + R_3(e^{i\theta}, tz) \quad \forall z \in V, t \in (0, t_1), e^{i\theta} \in \mathbb{T},$$

where the function R_3 is continuous on $\mathbb{T} \times t_1 V$ and

$$(5.18) \quad \sup_{e^{i\theta} \in \mathbb{T}} \frac{R_3(e^{i\theta}, tz)}{t^2 \|z\|_{\mathbb{C}^N}} \rightarrow 0 \quad \text{as } t \rightarrow 0.$$

Pick $h \in N(\partial \tilde{\Gamma} / \partial z)(P(\cdot)) \cap A_N$.

Since (f^*, α^*) solves UNCOPT, then there exists $t_1 > 0$ such that

$$(5.19) \quad \sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta}) + th(e^{i\theta})) \\ = \sup_{\theta} \tilde{\Gamma}(e^{i\theta}, \alpha(e^{i\theta}, th(e^{i\theta})), f^*(e^{i\theta}) + th(e^{i\theta})) > \gamma^* \\ \triangleq \sup_{\theta} \Gamma(e^{i\theta}, f^*(e^{i\theta})) \quad \forall t \in (0, t_1).$$

Combine (5.19), (I'), and expansion (5.17) to obtain that

$$(5.20) \quad \sup_{\theta} \left\{ 4 \operatorname{Re} z^t \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z + \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right)^\dagger \right. \\ \cdot \left(\frac{\partial \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \right) \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right) \\ \left. + \bar{z}^\dagger \frac{\partial^2 \Gamma}{\partial \bar{z} \partial z}(P(e^{i\theta})) z + 2 \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z^2}(P(e^{i\theta})) z \right\} \\ + \frac{R_2(e^{i\theta}, th(e^{i\theta}))}{t^2} > 0 \quad \forall t \in (0, t_1).$$

The necessity of (III') follows from letting $t \rightarrow 0$ in (5.20).

We now prove sufficiency of (I')–(III') with strict inequality. For this, assume that there exist $h \in A_N$ and a sequence $\{t_n\}$ of positive real numbers such that for all $n \in \mathbb{N}$

$$(5.21) \quad \sup_{\theta} \sup_{\alpha \in \mathcal{R}(e^{i\theta})} \tilde{\Gamma}(e^{i\theta}, \alpha, f^*(e^{i\theta}) + t_n h(e^{i\theta})) \leq \gamma^*.$$

From (5.21), we obtain

$$(5.22) \quad \tilde{\Gamma}(e^{i\theta}, \alpha(e^{i\theta}, t_n h(e^{i\theta})), f^*(e^{i\theta}) + t_n h(e^{i\theta})) < \gamma^* \quad \forall e^{i\theta} \in \mathbb{T}.$$

Now inequality (5.22) and expansion (5.17) imply that

$$(5.23) \quad 2 \operatorname{Re} \frac{\partial \tilde{\Gamma}}{\partial z}(P(e^{i\theta}))^\dagger h(e^{i\theta}) + o(t_n) < 0, \quad e^{i\theta} \in \mathbb{T}.$$

Letting $n \rightarrow \infty$ in (5.23), we obtain that

$$(5.24) \quad 2 \operatorname{Re} \frac{\partial \tilde{\Gamma}}{\partial z}(P(e^{i\theta}))^\dagger h(e^{i\theta}) \leq 0 \quad \forall e^{i\theta} \in \mathbb{T},$$

and Lemma 2.9 now implies that $h \in N(\partial \tilde{\Gamma} / \partial z(P(\cdot)))$. Combining again expansion (5.17) with inequality (5.22), we obtain

$$(5.25) \quad \begin{aligned} & 4 \operatorname{Re} h(e^{i\theta})^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha} \cdot \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0) h(e^{i\theta}) + \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0) z \right)^\dagger \\ & \cdot \left(\frac{\partial \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \right) \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0) h(e^{i\theta}) \right)^\dagger \\ & + 2 \overline{h(e^{i\theta})}^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \bar{z}}(P(e^{i\theta})) h(e^{i\theta}) + 2 \operatorname{Re} h(e^{i\theta})^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z^2}(P(e^{i\theta})) h(e^{i\theta}) \\ & + t_n R_3(e^{i\theta}, t_n h(e^{i\theta})) \leq 0 \quad \forall e^{i\theta} \in \mathbb{T}. \end{aligned}$$

Letting $n \rightarrow \infty$ in (5.25), we see that strict inequality in (III') cannot hold. This proves the theorem. \square

Proof of Theorem 5.6. We only need to prove (5.5). Note that there exists $t_1 > 0$ such that

$$(5.26) \quad \frac{\partial \tilde{\Gamma}}{\partial \alpha}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz) = 0 \quad \forall t \in (0, t_1), \quad e^{i\theta} \in \mathbb{T}.$$

Differentiating in (5.26) with respect to t we obtain

$$(5.27) \quad \begin{aligned} & \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz)^\dagger \cdot 2 \operatorname{Re} \left(\frac{\partial \alpha}{\partial z}(e^{i\theta}, tz)^\dagger z \right) \\ & + 2 \operatorname{Re} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz)^\dagger z = 0 \\ & \quad \forall t \in (0, t_1), \quad e^{i\theta} \in \mathbb{T}, \quad z \in V. \end{aligned}$$

Set $t = 0$ in (5.27) to obtain

$$\begin{aligned} & \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \cdot 2 \operatorname{Re} \left(\frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right) + 2 \operatorname{Re} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger z = 0 \\ & \quad \forall z \in V, \quad e^{i\theta} \in \mathbb{T}, \end{aligned}$$

i.e.,

$$(5.28) \quad 2 \operatorname{Re} \left(\frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right) = -\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta}))^{-1} 2 \operatorname{Re} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger z$$

$\forall z \in V, \quad e^{i\theta} \in \mathbb{T}.$

Therefore we obtain from (5.28)

$$(5.29) \quad \begin{aligned} & \left(2 \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \right) ((\alpha(e^{i\theta}), tz) - \alpha^*(e^{i\theta})) \\ &= -2 \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \cdot \frac{\partial^2 \tilde{\Gamma}}{\alpha^2}(P(e^{i\theta}))^{-1} 2 \operatorname{Re} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger z + R_3(e^{i\theta}, z) \\ &= -2 \bar{z}^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta}))^{-1} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger z \\ & \quad - 2 \operatorname{Re} \left(z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger z \right) + R_3(e^{i\theta}, z), \end{aligned}$$

where the error R_3 satisfies

$$(5.30) \quad \sup_{\theta} \frac{R_3(e^{i\theta}, z)}{\|z\|^2} \rightarrow 0 \quad \text{as } z \rightarrow 0.$$

Finally, the equality $\partial \tilde{\Gamma} / \partial \alpha(P(\cdot)) = 0$ and (5.29) and (5.6) imply that there exists $\eta > 0$ such that, if $\|h\|_\infty < \eta$, then

$$(5.31) \quad \begin{aligned} & \tilde{\Gamma}(e^{i\theta}, \alpha(e^{i\theta}, th(e^{i\theta})), f^*(e^{i\theta}) + th(e^{i\theta})) \\ &= \tilde{\Gamma}(P(e^{i\theta})) + \frac{t^2}{2} \left\{ \overline{h(e^{i\theta})}^\dagger \left(-\frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta}))^{-1} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \right. \right. \\ & \quad \left. \left. + \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial z}(P(e^{i\theta})) \right) h(e^{i\theta}) \right. \\ & \quad \left. + \operatorname{Re} h(e^{i\theta})^\dagger \left(-\frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta}))^{-1} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \right. \right. \\ & \quad \left. \left. + \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial \alpha}(P(e^{i\theta})) h(e^{i\theta}) \right) \right\} \\ & \quad + s(e^{i\theta}, th(e^{i\theta})) \quad \forall t \in (0, t), \quad e^{i\theta} \in \mathbb{T}, \end{aligned}$$

where $\sup_{\theta} (s(e^{i\theta}, th(e^{i\theta}))/t^2 \|h(e^{i\theta})\|_{\mathbb{C}^N}^2) \rightarrow 0$ as $t \rightarrow 0$ whenever $h(e^{i\theta}) \neq 0$. At this point, we proceed as in the proof of Theorem 5.7 to obtain the conclusion of the theorem.

Proof of Theorem 5.5. The first half of this theorem is already proved (Proof of Theorem 5.7). Assume that (I') and (II') hold. The fact that $\tilde{\Gamma}(e^{i\theta_0}, \alpha^*(e^{i\theta_0}), \cdot)$ is strictly convex near $z = f^*(e^{i\theta_0})$ merely says that

$$(5.32) \quad \bar{z} \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial z}(P(e^{i\theta_0}))z + \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z^2}(P(e^{i\theta_0}))z > 0 \quad \forall z \in \mathbb{C}^N \setminus \{0\}.$$

On the other hand, since (t^*, α^*) is a regular interior point and the problem considered is UNCOPT-S, $(\partial^2 \tilde{\Gamma} / \partial \alpha^2)(P(e^{i\theta}))$ takes values that are nonpositive definite matrices; hence

$$(5.33) \quad -2 \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) 2 \operatorname{Re} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))z \geq 0$$

$\forall z \in \mathbb{C}^N, \quad e^{i\theta} \in \mathbb{T}.$

The combination of (5.32) and (5.33) shows that (5.5) holds with strict inequality if $h(e^{i\theta_0}) \neq 0$.

Note that (5.32) holds for all $e^{i\theta} \in \mathbb{T}$ that are close enough to $e^{i\theta_0}$. If $h(e^{i\theta_0}) = 0$, then we can pick $e^{i\theta_1}$ near $e^{i\theta_0}$ such that $h(e^{i\theta_1}) \neq 0$. Using (5.32) for $e^{i\theta} = e^{i\theta_1}$, we obtain (5.5) with strict inequality. \square

Our next task is to simplify (III'). The simplification is that the test is written in terms of η and $\partial\eta/\partial z$ rather than in terms of $\alpha(e^{i\theta}, z)$ and $\partial\alpha/\partial z(e^{i\theta}, z)$. The advantage is that η is a real-valued scalar function, while α is a \mathbb{C}^N -valued function. Thus the corollary gives a substantial reduction in dimension of the computation.

COROLLARY 5.11. *If, in addition to the hypotheses of Theorem 5.7, we have that, for each $e^{i\theta} \in \mathbb{T}$, the matrix*

$$(5.34) \quad \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial^2 g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta}))$$

is invertible, then condition (III'), below, is equivalent to condition (III''):

$$(III'') \quad \sup_{\theta} \{ \overline{h(e^{i\theta})}^\dagger A_1(e^{i\theta}) h(e^{i\theta}) + \operatorname{Re} h(e^{i\theta})^\dagger B_1(e^{i\theta}) h(e^{i\theta}) \} \geq 0,$$

where the $N \times N$ matrix-valued functions $A_1(\cdot)$ and $B_1(\cdot)$ are defined by

$$\begin{aligned} A_1(\cdot) &\triangleq \left(\frac{\partial g}{\partial \alpha}(\cdot, \alpha^*(\cdot)) \frac{\partial \eta}{\partial z}(\cdot, 0) \right)^\dagger \left(\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(\cdot)) - \eta(\cdot, 0) \frac{\partial^2 g}{\partial \alpha^2}(\cdot, \alpha^*(\cdot)) \right)^{-1} \\ &\quad \cdot \left(\frac{\partial g}{\partial \alpha}(\cdot, \alpha^*(\cdot)) \frac{\partial \eta}{\partial z}(\cdot, 0) \right) - \left(\frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(\cdot)) \right) \\ &\quad \cdot \left(\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(\cdot)) - \eta(\cdot, 0) \frac{\partial^2 g}{\partial \alpha^2}(\cdot, \alpha^*(\cdot)) \right)^{-1} \left(\frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(\cdot)) \right)^\dagger + \frac{\partial^2 \tilde{\Gamma}}{\partial \bar{z} \partial z}(P(\cdot)) \\ B_1(\cdot) &\triangleq \left(\frac{\partial g}{\partial \alpha}(\cdot, \alpha^*(\cdot)) \frac{\partial \eta}{\partial z}(\cdot, 0) \right)^\dagger \left(\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(\cdot)) - \eta(\cdot, 0) \frac{\partial^2 g}{\partial \alpha^2}(\cdot, \alpha^*(\cdot)) \right)^{-1} \\ &\quad \cdot \left(\frac{\partial g}{\partial \alpha}(\cdot, \alpha^*(\cdot)) \frac{\partial \eta}{\partial z}(\cdot, 0) \right) - \left(\frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(\cdot, \cdot)) \right) \\ &\quad \cdot \left(\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(\cdot)) - \eta(\cdot, 0) \frac{\partial^2 g}{\partial \alpha^2}(\cdot, \alpha^*(\cdot)) \right)^{-1} \left(\frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(\cdot)) \right)^\dagger + \frac{\partial^2 \tilde{\Gamma}}{\partial z^2}(P(\cdot)). \end{aligned}$$

Proof. We have the relation

$$(5.35) \quad \frac{\partial \tilde{\Gamma}}{\partial \alpha}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz) - \eta(e^{i\theta}, tz) \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha(e^{i\theta}, tz)) = 0,$$

valid in some neighborhood of $0 \in \mathbb{C}^N$.

Differentiating with respect to t in (5.35), we have

$$\begin{aligned}
 (5.36) \quad & \frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz) 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, tz)^\dagger z \\
 & + 2 \operatorname{Re} \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(e^{i\theta}, \alpha(e^{i\theta}, tz), f^*(e^{i\theta}) + tz)^\dagger z \\
 & - \eta(e^{i\theta}, tz) \frac{\partial^2 g}{\partial \alpha^2}(e^{i\theta}, \alpha(e^{i\theta}, tz)) 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, tz)^\dagger z \\
 & - \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha(e^{i\theta}, tz)) \cdot 2 \operatorname{Re} \frac{\partial \eta}{\partial z}(e^{i\theta}, tz)^\dagger z = 0.
 \end{aligned}$$

Now take $t = 0$ in (5.36) to obtain

$$\begin{aligned}
 (5.37) \quad 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z &= \left(\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial^2 g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \right)^{-1} \\
 &\cdot 2 \operatorname{Re} \left(\frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha^*(e^{i\theta})) \frac{\partial \eta}{\partial z}(e^{i\theta}, 0)^\dagger - \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger \right) z.
 \end{aligned}$$

Let $L(e^{i\theta}, z)$ be defined by the equation

$$\begin{aligned}
 (5.38) \quad L(e^{i\theta}, z) &= 4 \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z + \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right)^\dagger \\
 &\cdot \left(\frac{\partial^2 \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \right) \left(2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \right).
 \end{aligned}$$

From (5.37) and (5.38), we have that

$$\begin{aligned}
 (5.39) \quad L(e^{i\theta}, z) &= 4 \operatorname{Re} z^\dagger \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z \\
 &+ 2 \operatorname{Re} z^\dagger \left(\frac{\partial \eta}{\partial z}(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha^*(e^{i\theta}))^\dagger - \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \right) \\
 &\cdot 2 \operatorname{Re} \frac{\partial \alpha}{\partial z}(e^{i\theta}, 0)^\dagger z.
 \end{aligned}$$

Equation (5.40) again implies that

$$\begin{aligned}
 (5.40) \quad L(e^{i\theta}, z) &= \left(2 \operatorname{Re} \frac{\partial \eta}{\partial z}(e^{i\theta}, 0) \frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha^*(e^{i\theta}))^\dagger + \frac{\partial^2 \tilde{\Gamma}}{\partial z \partial \alpha}(P(e^{i\theta})) \right) \\
 &\cdot \left(\frac{\partial \tilde{\Gamma}}{\partial \alpha^2}(P(e^{i\theta})) - \eta(e^{i\theta}, 0) \frac{\partial^2 g}{\partial \alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \right)^{-1} \\
 &\cdot 2 \operatorname{Re} \left(\frac{\partial g}{\partial \alpha}(e^{i\theta}, \alpha^*(e^{i\theta})) \frac{\partial \eta^\dagger}{\partial z}(e^{i\theta}, 0) - \frac{\partial^2 \Gamma}{\partial z \partial \alpha}(P(e^{i\theta}))^\dagger z \right).
 \end{aligned}$$

Now use the relation

$$(2 \operatorname{Re} z^\dagger b^\dagger) A (2 \operatorname{Re} bz) = 2 \bar{z}^\dagger \bar{b}^\dagger Abz + 2 \operatorname{Re} z^\dagger b^\dagger Abz,$$

in (5.40) to rewrite (III'). This gives the desired formulation of (III'). \square

All our theorems assume that there is a function $\alpha(e^{i\theta}, z)$ solving UNC and varying smoothly with parameters. It is important to have available conditions that automatically guarantee this. We give an example in the following proposition.

PROPOSITION 5.12. *Let (f^*, α^*) be a pair of admissible functions such that $\partial\Gamma/\partial z \times (\cdot, \alpha^*(\cdot), f^*(\cdot))$ never equals zero on \mathbb{T} . Suppose that, for each $e^{i\theta} \in \mathbb{T}$, $\alpha = \alpha^*(e^{i\theta})$ is the unique optimizer of $\tilde{\Gamma}(e^{i\theta}, \alpha, f^*(e^{i\theta}))$ over α satisfying $g(e^{i\theta}, \alpha) \leq 0$. Then either one of statements (a) and (b), below, is sufficient for the existence of a function $\alpha(\cdot, \cdot)$, which appears in Theorem 5.7:*

(a) *For each $e^{i\theta} \in \mathbb{T}$, $g(e^{i\theta}, \alpha^*(e^{i\theta})) < 0$, and the matrix $\partial^2\tilde{\Gamma}/\partial\alpha^2(e^{i\theta}, \alpha^*(e^{i\theta}), f^*(e^{i\theta}))$ is positive definite if the problem is UNCOPT-I, and negative definite if it is UNCOPT-S';*

(b) *For each $e^{i\theta} \in \mathbb{T}$, $g(e^{i\theta}, \alpha^*(e^{i\theta})) = 0$, there exists a real-valued function $\eta_1(e^{i\theta})$ such that $\partial\tilde{\Gamma}/\partial\alpha(\cdot, \alpha^*(\cdot), f^*(\cdot)) - \eta_1(\cdot)(\partial g/\partial\alpha)(\cdot, \alpha^*(\cdot)) = 0$. Let $M(e^{i\theta})$ be a real $K \times K$ matrix whose columns form a basis for the set of vectors in \mathbb{R}^k that are orthogonal to $\partial g/\partial\alpha(e^{i\theta}, \alpha^*(e^{i\theta}))$. Then*

$$M^\dagger(e^{i\theta}) \left(\frac{\partial^2\tilde{\Gamma}}{\partial\alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta}), f^*(e^{i\theta})) - \eta_1(e^{i\theta}) \frac{\partial^2 g}{\partial\alpha^2}(e^{i\theta}, \alpha^*(e^{i\theta})) \right) M(e^{i\theta})$$

is positive definite if the problem is UNCOPT-I, and negative definite if it is UNCOPT-S'.

Proof. (a) The implicit function theorem applied to the function $(\partial\Gamma/\partial\alpha)(e^{i\theta}, \alpha, f^*(e^{i\theta}) + z)$ at $z = 0$ and $\alpha = \alpha^*(e^{i\theta})$ implies that there exists a neighborhood $V(e^{i\theta})$ of $\{e^{i\theta}\} \times \{0\}$ in $\mathbb{T} \times \mathbb{C}^N$ and a $C^{(1)}$ function $\alpha : V(e^{i\theta}) \rightarrow \mathbb{R}^k$ such that

$$(5.41) \quad \frac{\partial\tilde{\Gamma}}{\partial\alpha}(e^{i\theta}, \alpha(e^{i\theta}, z), f^*(e^{i\theta}) + z) = 0, \quad (e^{i\theta}, z) \in V(e^{i\theta}).$$

We can pick $V(e^{i\theta})$ so small that $\partial^2\tilde{\Gamma}/\partial\alpha^2(e^{i\theta}, \alpha(e^{i\theta}, z), f^*(e^{i\theta}) + z)$ does not change signature for $z \in V$, i.e., $\alpha = \alpha(e^{i\theta}, z)$ is unique among those satisfying $g(e^{i\theta}, \alpha) \leq 0$, for every $z \in V(e^{i\theta})$ ("Mountain Pass Theorem"). Clearly, (i), (iii), and (iv) in Theorem 5.7 hold, and (ii) does hold, at least in a neighborhood of each $e^{i\theta} \in \mathbb{T}$. By a compactness argument, we see that (ii) also holds. This proves (a). The proof of (b) is similar to that of (a). \square

Appendix A. We have the following theorem.

THEOREM A1. *Let $q \in L_N^\infty$ be such that $\|q(\cdot)\|_{\mathbb{C}^N}$ is bounded away from 0 on \mathbb{T} . Suppose that $q \neq H_N^\infty$ and*

$$\inf_{f \in H^\infty} \|q - f\|_\infty < \inf\{\|q - g\|_\infty : g \text{ is continuous on } \mathbb{T}\}.$$

Then

- (a) *the function 0 is a best approximation to q from H_N^∞ if and only if*
 - (1) $\|q(e^{i\theta})\|_{\mathbb{C}^N}$ *is constant almost everywhere,*
 - (2) *the function $\chi^{-1}\bar{q}$ has an analytic direction F ;*
- (b) *moreover, if $q \in RL_N^1$, $g_0 \in RL_N^1$, then 1 and 2 with $F \in RH_N^1$ are necessary and sufficient for 0 to be a best approximation to q from RH_N^∞ .*

THEOREM A2. *Under the hypotheses of Theorem A1, if the set $\{e^{i\theta} \in \mathbb{T} : q \text{ is not continuous at } \theta\}^-$ has (linear) Lebesgue measure zero, then*

$$\inf_{f \in H_N^\infty} \|q - f\|_\infty = \inf_{f \in A_N} \|q - f\|_\infty.$$

The conclusion also holds if H_N^∞ and A_N are replaced by RH_N^∞ and RA_N , and if $q \in RL_N^1$, $g_0 \in RL_N^1$.

THEOREM A3. *Under the hypotheses of Theorem A1, suppose that 0 is a best approximation to q . Then*

- (1) $1 \leq \omega(\bar{q}) < \infty$,
 (2) $\chi^{-\omega(\bar{q})}\bar{q}$ has a unique analytic direction F_0 ,
 (3) Let $F \in H_N^1$ be such that $\|F\|_{L_N^1} = 1$. Then F is an analytic direction for \bar{q} if and only if there exist elements $z_1, \dots, z_{\omega(\bar{q})}$ in \mathbb{D}^- such that

$$F(e^{i\theta}) = \gamma \left(\prod_{l=1}^{\omega(\bar{q})} (e^{i\theta} - z_l)(1 - \bar{z}_l e^{i\theta}) \right) F_0(e^{i\theta}).$$

Here γ is a normalizing factor.

The proof of Theorems A1–A3 can be found in [M2].

Appendix B. The following lemma proves facts used in §3.

LEMMA B1. Let $\psi : \mathbb{R}^n \times K \rightarrow \mathbb{R}$, be of class C^1 , K compact in \mathbb{R}^p , be a function such that

$$\left(\frac{\partial \psi}{\partial x_1}(0, \kappa), \dots, \frac{\partial \psi}{\partial x_n}(0, \kappa) \right) \neq 0 \in \mathbb{R}^n \quad \forall \kappa \in K.$$

Then there exists an open neighborhood V of 0 in \mathbb{R}^n , such that, for all $k \in K$ and all $s \in \mathbb{R}$, the set

$$W(k, s) = \{x \in V : \psi(x, k) = s\}$$

is connected.

Proof. Fix $k_0 \in K$ and assume that $(\partial \psi / \partial x_1)(0, k_0) \neq 0$. Define a function $\varphi : \mathbb{R}^n \times k \rightarrow \mathbb{R}^n \times K$ by $\varphi((x_1, \dots, x_n), k) = ((\psi((x_1, \dots, x_n), k), x_2, \dots, x_n), k)$; hence the Jacobian

$$D\varphi(0, k_0) = \begin{pmatrix} \frac{\partial \psi}{\partial x_1}(0, k_0) & \frac{\partial \psi}{\partial x_2}(0, k_0) & \cdots & \frac{\partial \psi}{\partial x_n}(0, k_0) & \frac{\partial \psi}{\partial k}(0, k_0) \\ 0 & 1 & & \cdot & \cdot \\ 0 & 0 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & 1 & \\ 0 & 0 & \cdots & & 1 \end{pmatrix}$$

is invertible, and therefore there exists neighborhoods V_{k_0} of 0 in \mathbb{R}^n , U_{k_0} of k_0 in K such that φ is a diffeomorphism on $V_{k_0} \times U_{k_0}$. Now, if $((x_1, \dots, x_n), k) = \varphi^{-1}((y_1, \dots, y_n), k)$, then

$$(B.1) \quad y_1 = \psi((x_1, \dots, x_n), k) = \psi \varphi^{-1}((y_1, \dots, y_n), k),$$

and this shows that, for each $k \in U_{k_0}$ and each $z \in \mathbb{R}$, the set

$$A_i(z) \triangleq \{((y_1, \dots, y_n), k) \in \varphi(V_{k_0} \times U_{k_0}) : \psi \varphi^{-1}((y_1, \dots, y_n), k) = z\}$$

is connected. Hence also connected is the set

$$\varphi^{-1}(A_k(z)) = \{((x_1, \dots, x_n), k) \in V_{k_0} \times U_{k_0} : \psi((x_1, \dots, x_n), k) = z\}.$$

Finally, the sets $U_k, k \in K$, cover K , so there exists a finite cover $\{U_{k_i}\}_{i=1}^p$, and, to complete the proof, we can set $V = \bigcap_{i=1}^p V_{k_i}$. \square

REFERENCES

- [A] L. V. AHLFORS, *Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1966.
- [BB] C. C. BARRAT AND S. P. BOYD, *Linear Controller Design*, Prentice-Hall, 1991.
- [BGR] J. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, Operator Theory: Advances and Applications Series, Vol. 45, Birkhäuser, Boston, 1991.
- [BHM] J. BENCE, J. W. HELTON, AND D. MARSHALL, H^∞ optimization, in Proc. Conference on Decision and Control, Athens, 1986.
- [Dor] P. DORATO AND R. K. YEDAVALLI, eds., *Recent Advances in Robust Control*, IEEE Press, New York, 1990.
- [Doy] J. C. DOYLE, *Synthesis of robust controllers*, in Proc. IEEE Conference on Decision and Control, December 1983, San Antonio, TX, pp. 109–114.
- [Dym] H. DYM, *J Contractive Matrix Functions Reproducing Kernel Hilbert Spaces and Interpolation*, in Regional Conferences Series in Mathematics, 71, Cleveland, OH, 1989.
- [FF] C. FOIAS AND A. FRAZHO, *The commutant lifting approach to interpolation problems*, in Operator Theory: Advances and Applications, Vol. 0T44, 1990.
- [Fr] B. A. FRANCIS, *A Course in H^∞ Control Theory*, Lecture Notes in Control and Information Sci., Vol. 88, Springer-Verlag, New York, Berlin, 1986.
- [G] J. GARNETT, *Bounded Analytic Functions*, Academic Press, New York, 1981.
- [GM] K. GLOVER AND D. C. MCFARLANE, *Robust controller design using normalized coprime factor plant descriptions*, in Lecture Notes in Control and Information Sciences, 138, Springer-Verlag, New York, Berlin, 1990.
- [GMW] P. GILL, W. MURRAY, AND M. WRIGHT, *Practical Optimization*, Academic Press, London, 1984.
- [Go] I. C. GOHBERG, *A factorization problem in normed rings, functions of isometric and symmetric operators, and singular integral equations*, Russian Math. Surveys, 19 (1964), pp. 63–114.
- [H1] J. W. HELTON, *Worst case analysis in the frequency domain: An H^∞ approach to control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1154–1170.
- [H2] ———, *Optimal frequency domain design vs. an area of several complex variables*, preprint.
- [H3] ———, *Optimization over spaces of analytic functions and the Corona problem*, J. Oper. Theory, 13 (1986), pp. 359–375.
- [H4] ———, *Operator Theory, Analytic Functions, Matrices, and Electrical Engineering*, in Regional Conference Series in Mathematics, No. 68, Lincoln, NE, August 1985, American Mathematical Society, Providence, RI, 1987.
- [H5] ———, *Non-Euclidean functional analysis and electronics*, Bull. Amer. Math. Soc., 7 (1982), pp. 1–64.
- [HH] J. W. HELTON AND R. HOWE, *A bang-bang principle for the frequency domain*, J. Approx. Theory, 47 (1986), pp. 101–121.
- [HM] J. W. HELTON AND D. MARSHALL, *Frequency domain design and analytic selections*, Indiana J. Math., 39 (1990).
- [HMer] J. W. HELTON AND O. MERINO, *Numerical results in H^∞ control*, in Proc. Amer. Control Conference, San Diego, CA, 1990.
- [HS] J. W. HELTON AND D. SCHWARTZ, *Analytic best approximation to vector valued functions*, preprint.
- [M1] O. MERINO, *Stability of qualitative properties and continuity of solutions to problems of optimization over spaces of analytic functions*, American J. Math., to appear.
- [M2] ———, *Optimizing Real Valued Linear Functionals on H^1* , unpublished.
- [RR1] M. ROSENBLUM AND J. ROVNYAK, *Hardy Classes and Operator Theory*, Oxford University Press, Oxford, UK, 1985.
- [RR2] ———, *The factorization problem for nonnegative operator valued functions*, Bull. Amer. Math. Soc., 77 (1971), pp. 287–318.
- [Yng] N. YOUNG, *An Introduction to Hilbert Space*, Cambridge University Press, Cambridge, UK, 1988.
- [YS] D. C. YOULA AND M. SAITO, *Interpolation with positive real functions*, J. Franklin Inst., 284 (1967), pp. 77–108.

CONTROLLABILITY AND STABILIZABILITY OF COUPLED STRINGS WITH CONTROL APPLIED AT THE COUPLED POINTS*

L. F. HO†

Abstract. Controllability and stabilizability of a system of coupled strings with control applied at the coupled points is studied. By investigating the properties of certain exponential series, it is shown that the system is approximate controllable if and only if related systems of uncoupled strings do not share a common eigenvalue. A sufficient condition for exact controllability is also obtained in terms of the Riesz basis properties of those exponential series.

Key words. coupled strings, controllability, stabilizability, nonharmonic Fourier series, Riesz basis

AMS subject classification. 93B05

1. Introduction. Let $L > 0$ and $0 = x_0 < x_1 < \cdots < x_N = L$. We consider the system of coupled strings

$$(1.1) \quad \frac{\partial^2 y}{\partial t^2} = \frac{1}{\rho_i} \frac{\partial}{\partial x} \left(\tau_i \frac{\partial y}{\partial x} \right), \quad x \in (x_{i-1}, x_i),$$

with conditions at the boundary and at the coupled points

$$(1.2a) \quad y(0, t) = y(L, t) = 0,$$

$$(1.2b) \quad y(x_i^+, t) = y(x_i^-, t), \quad i = 1, \dots, N-1,$$

and control applied at the coupled points

$$(1.3) \quad \tau_{i+1}(x_i) \frac{\partial y(x_i^+, t)}{\partial x} - \tau_i(x_i) \frac{\partial y(x_i^-, t)}{\partial x} = u_i(t), \quad i = 1, \dots, N-1,$$

where $u_i \in L^2[0, T]$, $i = 1, \dots, N-1$, and the functions $\rho_i, \tau_i, i = 1, \dots, N$ are continuously differentiable and positive. We will consider the state space $X = H_0^1[0, L] \times L^2[0, L]$ and the space of controls $U = (L^2[0, T])^{N-1}$. We say that the system is *exactly controllable in time* $T > 0$ if given any initial state (f_1, f_2) and terminal state (g_1, g_2) , both in X , there exist controls (u_1, \dots, u_{N-1}) in U such that the corresponding solution of (1.1) with initial conditions

$$(1.4) \quad y(x, 0) = f_1(x); \quad \frac{\partial}{\partial t} y(x, 0) = f_2(x)$$

satisfies

$$y(x, T) = g_1(x); \quad \frac{\partial}{\partial t} y(x, T) = g_2(x).$$

We say that the system is *approximately controllable in time* $T, T > 0$, if given any initial state (f_1, f_2) and terminal state (g_1, g_2) both in X and any $\epsilon > 0$, there exist controls (u_1, \dots, u_{N-1}) in U such that the solution of (1.1) with initial condition (1.4) satisfies

$$\left\| \left(y(\cdot, T), \frac{\partial y(\cdot, T)}{\partial t} \right) - (g_1, g_2) \right\|_X < \epsilon.$$

* Received by the editors April 30, 1990; accepted for publication (in revised form) April 8, 1992.

† Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435.

We also consider the related problem of stabilizability with feedback given by

$$u_i(x, t) = k_i \frac{\partial y(x_i, t)}{\partial t}, \quad i = 1, \dots, N-1,$$

where $k_i, i = 1, \dots, N-1$, are positive constants. Controllability of the system will then be established by the “controllability via stabilizability” method of Russell [15]. The stabilizability problem, for the case of constant wave speeds, has been considered by Chen, Coleman, and West [1], Liu [11], and Liu, Huang, and Chen [12]. There is some similarity between their results and the result we obtain in this paper. We will discuss that in the examples given in §5.

In §2, we will establish some results concerning exponential series that we will need for the proof of the main results. In §3, we will state and prove our main results on stabilizability and approximate controllability. Roughly speaking, it says that the whole system is approximately controllable if and only if the uncoupled strings do not share a common eigenvalue. In §4, we consider stabilizability with a uniform exponential decay rate and the related problem of exact controllability. A sufficient condition is given in terms of the Riesz basis property for certain sets of exponential functions. Some examples are discussed in §5.

2. Nonharmonic Fourier series. In this section, we study sequences of exponential functions of the form $\{e^{i\lambda_n t}\}_{n=-\infty}^{\infty}$ where $\lambda_n, -\infty < n < \infty$, are real. Properties of completeness and independence of series of the form $\sum_n a_n e^{i\lambda_n t}$, known as nonharmonic Fourier series, have been investigated by, e.g., Levinson [8] and Riesz and Nagy [13]. Russell has also studied the relationship between the properties of such series and results in control theory [14], [15]. We first recall a definition.

DEFINITION 2.1. Let X be a Hilbert space and let $\{\phi_n\}_{n=-\infty}^{\infty}$ be a sequence in X . We say that $\{\phi_n\}_{n=-\infty}^{\infty}$ is a *Riesz basis* of X if

(RB1) $\overline{\text{span}}\{\phi_n : -\infty < n < \infty\} = X$;

(RB2) there exist constants $d, D > 0$ such that

$$(2.1) \quad d \sum_n |a_n|^2 \leq \left\| \sum_n a_n \phi_n \right\|^2 \leq D \sum_n |a_n|^2$$

for any finite sequence $\{a_n\}_{n=-\infty}^{\infty}$.

If a sequence $\{\phi_n\}_{n=-\infty}^{\infty}$ satisfies the first inequality in (RB2), we say that it is *uniformly independent*. We also have the following weaker sense of independence. (See Levinson and McCalla [9].)

DEFINITION 2.2. Let $\{\phi_n\}_{n=-\infty}^{\infty}$ be a sequence in a Hilbert space X . Suppose for all n there exists a positive constant d_n such that

$$(2.2) \quad \left\| \sum_k a_k \phi_k \right\| \geq d_n |a_n|$$

for all finite sequences $\{a_n\}_{n=-\infty}^{\infty}$; then we say that $\{\phi_n\}_{n=-\infty}^{\infty}$ is *strongly independent* in X .

The following fact is well known.

A sequence $\{\phi_n\}_{n=-\infty}^{\infty}$ is strongly independent if and only if we can find a sequence $\{\psi_n\}_{n=-\infty}^{\infty}$ such that

$$[\psi_m, \phi_n] = \delta_{m,n}.$$

The sequence $\{\psi_n\}_{n=-\infty}^{\infty}$ is said to be *biorthogonal* to $\{\phi_n\}_{n=-\infty}^{\infty}$.

If $f : \mathbb{R} \rightarrow \mathbb{C}$ has compact support, we denote

$$\hat{f}(\lambda) = \int_{-\infty}^{\infty} f(t)e^{i\lambda t} dt.$$

Obviously, \hat{f} is an entire function of λ .

Let $a < b$ and f be any function in $L^2[a, b]$; we may consider f as defined on the whole real line by extending it by zero outside $[a, b]$. In the rest of this paper, we will identify any such function f with its extension in this way.

LEMMA 2.3. *Let $f \in L^2[0, T]$ and let μ be a zero of \hat{f} of order greater than or equal to m , then*

$$\frac{\hat{f}(\lambda)}{(\lambda - \mu)^m} = \hat{g}(\lambda)$$

for some function g in $L^2[0, T]$.

Proof. We prove by induction on m . For $m = 1$, we make the following straightforward calculation:

$$\begin{aligned} \frac{\hat{f}(\lambda)}{\lambda - \mu} &= \frac{\hat{f}(\lambda) - \hat{f}(\mu)}{\lambda - \mu} = \int_0^T f(t) \frac{e^{i\lambda t} - e^{i\mu t}}{\lambda - \mu} dt \\ &= \int_0^T f(t) i \int_0^t e^{i\lambda s + i\mu(t-s)} ds dt \\ &= \int_0^T i \int_s^T f(t) e^{i\mu(t-s)} dt e^{i\lambda s} ds \\ &= \int_0^T g(t) e^{i\lambda t} dt = \hat{g}(\lambda), \end{aligned}$$

where $g(t) = i \int_t^T f(s) e^{i\mu(s-t)} ds$.

Suppose the result holds for $m \leq k$. Let μ be a zero of order greater than or equal to $k + 1$ of \hat{f} . By induction hypothesis,

$$\frac{\hat{f}(\lambda)}{(\lambda - \mu)^k} = \hat{h}(\lambda)$$

for some h in $L^2[0, T]$. Then μ is a zero of order greater than or equal to 1 of \hat{h} . By induction hypothesis once again, we have

$$\frac{\hat{f}(\lambda)}{(\lambda - \mu)^{k+1}} = \frac{\hat{h}(\lambda)}{\lambda - \mu} = \hat{g}(\lambda)$$

for some g in $L^2[0, T]$. This completes the proof of the lemma.

LEMMA 2.4. *If $\{e^{i\lambda_n t} : -\infty < n < \infty\}$ is strongly independent in $L^2[0, T]$, then for any real number ξ and any $T_1 > T$, the set*

$$\{e^{i\lambda_n t} : -\infty < n < \infty\} \cup \{e^{i\xi t}\}$$

is strongly independent in $L^2[0, T_1]$.

Proof. If $\xi = \lambda_n$ for some n , the result is trivial. So let us assume that $\xi \neq \lambda_n$ for all n . Let $\{p_n\}_{n=-\infty}^{\infty}$ be functions in $L^2[0, T]$ biorthogonal to $\{e^{i\lambda_n t} : -\infty < n < \infty\}$. Let $\delta = T_1 - T$ and let $\{g_n : -\infty < n < \infty\}$ and h be functions in $L^2[0, \delta]$ such that

$$\int_0^\delta g_n(t) e^{i\lambda_n t} dt = 1,$$

$$\int_0^\delta g_n(t) e^{i\xi t} dt = 0, \quad -\infty < n < \infty,$$

and

$$\int_0^\delta h(t) e^{i\lambda_0 t} dt = 0.$$

Let $q_n = p_n^* g_n$. (Here $*$ denotes convolution). Then q_n is supported on $[0, T_1]$ and $\hat{q}_n(\lambda) = \hat{p}_n(\lambda) \hat{g}_n(\lambda)$. So we have

$$\int_0^{T_1} q_n e^{i\lambda_m t} dt = \left(\int_0^T p_n e^{i\lambda_m t} dt \right) \left(\int_0^\delta g_n e^{i\lambda_m t} dt \right) = \delta_{m,n},$$

$$\int_0^{T_1} q_n e^{i\xi t} dt = \left(\int_0^T p_n e^{i\xi t} dt \right) \left(\int_0^\delta g_n e^{i\xi t} dt \right) = 0.$$

Let $r_0 = p_0^* h$. Then again r_0 is supported on $[0, T]$ and $\hat{r}_0(\lambda) = \hat{p}_0(\lambda) \hat{h}(\lambda)$. Let ξ be a zero of order m of \hat{r}_0 (m may be zero). By Lemma 2.3, there exists r in $L^2[0, T_1]$ such that $\hat{r}(\lambda) = \hat{r}_0(\lambda)/(\lambda - \xi)^m$. Since $\hat{r}_0(\lambda_n) = 0$ for all n , we have

$$\int_0^{T_1} r(t) e^{i\lambda_n t} dt = \hat{r}(\lambda_n) = 0 \quad \text{for all } n.$$

Because $\hat{r}_0^{(m)}(\xi) \neq 0$,

$$\int_0^{T_1} r(t) e^{i\xi t} dt = \hat{r}(\xi) = \hat{r}_0^{(m)}(\xi)/m! \neq 0.$$

It follows that $\{q_n : -\infty < n < \infty\} \cup \{m! r / \hat{r}_0^{(m)}(\xi)\}$ is biorthogonal to $\{e^{i\lambda_n t} : -\infty < n < \infty\} \cup \{e^{i\xi t}\}$ in $L^2[0, T_1]$. Hence the latter is strongly independent in $L^2[0, T_1]$.

Remark. Clearly, Lemma 2.4 remains valid if the set $\{e^{i\xi t}\}$ is replaced by any finite set $\{e^{i\xi_n t} : 1 \leq n \leq m\}$.

THEOREM 2.5. Suppose that $\{e^{i\lambda_n t}\}_{n=-\infty}^{\infty}$ and $\{e^{i\mu_n t}\}_{n=-\infty}^{\infty}$ are strongly independent in $L^2[0, T_1]$ and $L^2[0, T_2]$, respectively. Then for any $T > T_1 + T_2$, the set

$$S = \{e^{i\lambda_n t} : -\infty < n < \infty\} \cup \{e^{i\mu_n t} : -\infty < n < \infty\}$$

is strongly independent in $L^2[0, T]$.

Proof. It suffices to show the existence of a sequence biorthogonal to S . Let the sequences $\{p_n\}_{n=-\infty}^{\infty}$ and $\{q_n\}_{n=-\infty}^{\infty}$ be biorthogonal to $\{e^{i\lambda_n t}\}_{n=-\infty}^{\infty}$ and $\{e^{i\mu_n t}\}_{n=-\infty}^{\infty}$, respectively. Let $e^{i\eta \cdot} \in S$. We need to find a square integrable function r vanishing outside $[0, T]$ such that

$$(2.3) \quad \int_0^T r(t) e^{i\xi t} dt = \hat{r}(\xi) = 0 \quad \text{if } e^{i\xi \cdot} \in S \quad \text{and} \quad \xi \neq \eta$$

and

$$(2.4) \quad \int_0^T r(t)e^{i\eta t} dt = \hat{r}(\eta) = 1.$$

We consider three cases.

Case 1. $\eta = \lambda_n = \mu_m$ for some m and n .

Let $r = p_n^* q_m$. (Here $*$ again denotes convolution.) Then r vanishes outside $[0, T]$ and $\hat{r}(\lambda) = \hat{p}_n(\lambda)\hat{q}_m(\lambda)$ for any complex number λ . It is easy to see that (2.3) and (2.4) hold.

Case 2. $\eta = \lambda_n$ but does not equal any μ_m . Since $T - T_1 > T_2$, by Lemma 2.4, the set $\{e^{i\mu_m t} : -\infty < m < \infty\} \cup \{e^{i\eta t}\}$ is strongly independent in $L^2[0, T - T_1]$. In particular, we can find a function f in $L^2[0, T - T_1]$ such that $\hat{f}(\mu_m) = 0$ for all m and $\hat{f}(\eta) = 1$. Let $r = p_n^* f$. Then r is in $L^2[0, T]$ and (2.3) and (2.4) hold.

Case 3. $\eta = \mu_m$ for some m but not equal to any λ_n . We can find r in a way similar to that in Case 2.

This completes the proof of the theorem.

It follows immediately from the remark after Lemma 2.4 that Theorem 2.5 can be strengthened slightly as follows.

THEOREM 2.6. *Suppose $\{e^{i\lambda_n t}\}_{n=-\infty}^{\infty}$ and $\{e^{i\mu_n t}\}_{n=-\infty}^{\infty}$ are strongly independent in $L^2[0, T_1]$ and $L^2[0, T_2]$, respectively, and $\{\xi_n : 1 \leq n \leq m\}$ is any finite set of real numbers. Then for any $T > T_1 + T_2$, the set*

$$S = \{e^{i\lambda_n t} : -\infty < n < \infty\} \cup \{e^{i\mu_n t} : -\infty < n < \infty\} \cup \{e^{i\xi_n t} : 1 \leq n \leq m\}$$

is strongly independent in $L^2[0, T]$.

3. Approximate controllability. We first consider the stability of the following closed-loop system:

$$(3.1) \quad \frac{\partial^2 y}{\partial t^2} = \frac{1}{\rho_i} \frac{\partial}{\partial x} \left(\tau_i \frac{\partial y}{\partial x} \right), \quad x \in (x_{i-1}, x_i);$$

$$(3.2a) \quad y(0, t) = y(L, t) = 0,$$

$$(3.2b) \quad y(x_i^+, t) = y(x_i^-, t),$$

$$(3.3) \quad \tau_{i+1}(x_i) \frac{\partial y(x_i^+, t)}{\partial x} - \tau_i(x_i) \frac{\partial y(x_i^-, t)}{\partial x} = k_i \frac{\partial y(x_i, t)}{\partial t}, \quad i = 1, \dots, N-1,$$

$$(3.4) \quad y(x, 0) = f_1(x), \quad \frac{\partial y(x, 0)}{\partial t} = f_2(x),$$

where $k_i, i = 1, \dots, N-1$, are positive constants.

We will show that the stability of the above system is related to the properties of a system of N uncoupled strings.

$$(3.5) \quad \frac{\partial^2 w}{\partial t^2} = \frac{1}{\rho_i} \frac{\partial}{\partial x} \left(\tau_i \frac{\partial w}{\partial x} \right), \quad x \in (x_{i-1}, x_i);$$

$$(3.6) \quad w(x_{i-1}, t) = w(x_i, t) = 0,$$

$i = 1, \dots, N$. We will refer to the i th string in this system as $(S_i), i = 1, \dots, N$. Recall that the functions $\rho_i, \tau_i, i = 1, \dots, N$, are assumed to be continuously differentiable and positive. We first give a result of strong observability for such a system. When the wave speed is constant or when the domain and the wave speeds are both analytic, this result is known, even in higher dimensions. (See Bardos, Lebeau, and Rauch [1], Ho [4], [5], and Kormonik [7]).

THEOREM 3.1. *The systems $(S_i), i = 1, \dots, N - 1$ are strongly observable in some time $T_i > 0$ with observation*

$$(3.7) \quad \theta_i(t) = \frac{\partial}{\partial x} w(x_{i-1}, t)$$

or

$$(3.8) \quad \theta_i(t) = \frac{\partial}{\partial x} w(x_i, t).$$

In other words, for $i = 1, \dots, N$ the inequality

$$\int_0^T \theta_i(t)^2 dt \geq K_i \int_0^L \tau_i \left(\frac{\partial}{\partial x} w(x, 0) \right)^2 + \rho_i \left(\frac{\partial}{\partial t} w(x, 0) \right)^2$$

holds for some positive constants K_i .

Proof. We use a multiplier method similar to that used in Ho [6]. For simplicity of notation, let us drop the subscript i and assume that the string extends from $x = 0$ to $x = L$. We first consider the observation

$$(3.9) \quad \theta(t) = \frac{\partial}{\partial x} w(L, t).$$

Let h be the function satisfying the linear differential equation

$$(3.10) \quad h'(x) = 1 + \max \left\{ -\frac{\rho'}{\rho}, \frac{\tau'}{\tau} \right\} h(x)$$

with initial condition

$$(3.11) \quad h(0) = 0.$$

It is clear that $h(x) > 0$ for $x > 0$. Multiplying (3.1) by $\rho h(\partial w / \partial x)$ and integrating, we have

$$\int_0^T \int_0^L h \frac{\partial w}{\partial x} \left(\rho \frac{\partial^2 w}{\partial t^2} - \frac{\partial}{\partial x} \left(\tau \frac{\partial w}{\partial x} \right) \right) dx dt = 0.$$

Integrating by parts and making use of the boundary conditions for w and h , we have

$$\begin{aligned} \frac{1}{2} \int_0^T \int_0^L \frac{d}{dx} (h\rho) \left(\frac{\partial w}{\partial t} \right)^2 + \frac{d}{dx} \left(\frac{h}{\tau} \right) \left(\tau \frac{\partial w}{\partial x} \right)^2 dx dt + \int_0^L h\rho \frac{\partial w}{\partial x} \frac{\partial w}{\partial t} dx \Big|_{t=0}^{t=T} \\ - \frac{1}{2} h(L)\tau(L) \int_0^T \theta(t)^2 dt = 0. \end{aligned}$$

Then by the differential equation (3.9) and conservation of energy, we have

$$(3.12) \quad \frac{1}{2}h(L)\tau(L) \int_0^T \theta(t)^2 dt \leq TE_0 - KE_0,$$

where

$$E_0 = \frac{1}{2} \int_0^L \tau \left(\frac{\partial}{\partial x} w(x, 0) \right)^2 + \rho \left(\frac{\partial}{\partial t} w(x, 0) \right)^2 dx$$

and K is a constant that only depends on h and ρ . It follows from (3.12) that the system (3.5), (3.6) with observation (3.9) is strongly observable in any T such that $T > K$. The proof for strong observability for the observation $\theta(t) = (\partial/\partial x)w(0, t)$ is similar.

We let $\{\lambda_{n,i}\}_{n=1}^\infty$ be the sequence of eigenvalues of the system (S_i) and $\{\phi_{n,i}\}_{n=1}^\infty$ be the corresponding set of eigenfunctions satisfying

$$(3.13) \quad \frac{1}{\rho_i} \frac{\partial}{\partial x} \left(\tau_i \frac{\partial}{\partial x} \phi_{n,i} \right) + \lambda_{n,i}^2 \phi_{n,i} = 0, \quad \phi_{n,i}(x_{i-1}) = \phi_{n,i}(x_i) = 0,$$

and we assume that $\phi_{n,i}$ is normalized so that

$$(3.14) \quad \int_{x_{i-1}}^{x_i} \tau_i \left(\frac{\partial}{\partial x} \phi_{n,i} \right)^2 dx = \int_{x_{i-1}}^{x_i} \lambda_{n,i}^2 \rho_i \phi_{n,i}^2 dx = 1.$$

It is well known that strong observability is related to the Riesz basis property (more precisely, the property (RB2) in §1) of the exponential functions formed from the eigenvalues of the system (see Russell [14], [15]). The following proposition follows easily from known results but we supply its proof for completeness.

PROPOSITION 3.2. *If the system (S_i) is strongly observable in time T_i with observation $\theta_i(t)$ given by (3.8) or (3.9), then the sequence of exponential functions $\{e^{\pm i\lambda_{n,i}t}\}_{n=1}^\infty$ satisfies condition (RB2) in the space $L^2[0, T_i]$.*

Proof. Suppose the system is strongly observable in time T_i with observation $\theta_i(t)$ given by (3.9), we then have the inequality

$$(3.15) \quad \int_0^{T_i} |\theta_i(t)|^2 dt \geq K_1 E_0,$$

where

$$\begin{aligned} E_0 &= \text{total energy of } w \text{ at } t = 0 \\ &= \frac{1}{2} \int_{x_{i-1}}^{x_i} \rho_i \left(\frac{\partial w(x, 0)}{\partial t} \right)^2 + \tau_i \left(\frac{\partial w(x, 0)}{\partial x} \right)^2 dx, \end{aligned}$$

w is any solution of (3.5), (3.6), and K_1 is a constant independent of w . Also by regularity results (Lasiecka, Lions, and Triggiani [10]), there exists a constant $K_2 > 0$ such that

$$(3.16) \quad \int_0^{T_i} |\theta_i(t)|^2 dt \leq K_2 E_0.$$

Because every function $e^{i\lambda_{n,i}t} \phi_{n,i}(x)$, $1 \leq n < \infty$, is a solution of (3.5), (3.6) and by the normalization condition (3.13), the total energy $E_0 = 1$, it follows that

$$(3.17) \quad K_1 \leq \int_0^{T_i} |e^{i\lambda_{n,i}t} \phi'_{n,i}(x_{i-1})|^2 dt \leq K_2.$$

Hence

$$(3.18) \quad K_1/T_i \leq |\phi'_{n,i}(x_{i-1})|^2 \leq K_2/T_i.$$

Let $\{a_n\}_{n=-\infty, n \neq 0}^\infty$ by a finite sequence and let $b_{\pm n} = a_{\pm n}/\phi'_{n,i}(x_{i-1})$, $n = 1, 2, \dots$. The function

$$(3.19) \quad w(x, t) = \sum_{n=1}^{\infty} (b_n e^{i\lambda_{n,i}t} + b_{-n} e^{-i\lambda_{n,i}t}) \phi_{n,i}(x)$$

is a solution of (3.5), (3.6) with energy

$$(3.20) \quad E_0 = \sum_{n=1}^{\infty} (|b_n|^2 + |b_{-n}|^2)$$

and observation

$$(3.21) \quad \theta_i(t) = \sum_{n=1}^{\infty} a_n e^{i\lambda_{n,i}t} + a_{-n} e^{-i\lambda_{n,i}t}.$$

By (3.18) and (3.20), we have

$$(3.22) \quad (T_i/K_2) \sum_{n=1}^{\infty} |a_n|^2 + |a_{-n}|^2 \leq E_0 \leq (T_i/K_1) \sum_{n=1}^{\infty} |a_n|^2 + |a_{-n}|^2.$$

Hence by (3.15) and (3.16), we have

$$\begin{aligned} \frac{T_i K_1}{K_2} \sum_{n=1}^{\infty} |a_n|^2 + |a_{-n}|^2 &\leq \int_0^{T_i} \left| \sum_{n=1}^{\infty} a_n e^{i\lambda_{n,i}t} + a_{-n} e^{-i\lambda_{n,i}t} \right|^2 dt \\ &\leq \frac{T_i K_2}{K_1} \sum_{n=1}^{\infty} |a_n|^2 + |a_{-n}|^2. \end{aligned}$$

This completes the proof of the proposition when θ_i is given by (3.9). The proof for θ_i given by (3.8) is similar.

Let us now consider the system (3.1)–(3.3). Let $T > 0$. Given any initial state $(f_1, f_2) \in H_0^1[0, L] \times L^2[0, L]$, we can find the solution of (3.1)–(3.3) with initial conditions (3.4). Denote

$$\begin{aligned} \mathcal{E}y(t) &= \text{the total energy of } y \text{ at time } t \\ (3.23) \quad &= \frac{1}{2} \sum_{k=1}^{N-1} \int_{x_i}^{x_{i+1}} \rho_i(x) \left(\frac{\partial y(x, t)}{\partial t} \right)^2 = \tau_i(x) \left(\frac{\partial y(x, t)}{\partial x} \right)^2 dx. \end{aligned}$$

We have

$$\begin{aligned} (3.24) \quad \frac{d\mathcal{E}y(t)}{dt} &= \sum_{k=1}^{N-1} \frac{\partial y(x_i, t)}{\partial t} \left(\tau_i(x_i) \frac{\partial y(x_i^-, t)}{\partial x} - \tau_{i+1}(x_i) \frac{\partial y(x_i^+, t)}{\partial x} \right) \\ &= - \sum_{k=1}^{N-1} k_i \left(\frac{\partial y(x_i, t)}{\partial t} \right)^2 \end{aligned}$$

It follows that for any $t_1, t_2 > 0$, we have

$$(3.25) \quad \mathcal{E}y(t_1) - \mathcal{E}y(t_2) = - \sum_{k=1}^{N-1} \int_{t_1}^{t_2} k_i \left(\frac{\partial y(x_i, t)}{\partial t} \right)^2 dt.$$

Hence in particular,

$$\begin{aligned} \sum_{k=1}^{N-1} \int_0^T k_i \left(\frac{\partial y(x_i, t)}{\partial t} \right)^2 dt &= \mathcal{E}y(0) - \mathcal{E}y(T) \leq \mathcal{E}y(0) \\ &= \frac{1}{2} \sum_{k=1}^{N-1} \int_{x_i}^{x_{i+1}} \rho_i(x) f_2(x)^2 + \tau_i(x) f_1'(x)^2 dx \\ &\leq K \|f_1\|_{H_0^1[0, L]}^2 + \|f_2\|_{L^2[0, L]}^2, \end{aligned}$$

where K is a positive constant. So if we define

$$(3.26) \quad \mathcal{N}_T(f_1, f_2) = \left(\sum_{k=1}^{N-1} \int_0^T k_i \left(\frac{\partial y(x_i, t)}{\partial t} \right)^2 dt \right)^{1/2},$$

then $\mathcal{N}_T(\cdot, \cdot)$ is a seminorm on $H^1[0, L] \times L^2[0, L]$ and by (3.22), we have

$$(3.27) \quad \mathcal{E}y(t) - \mathcal{E}y(t+T) = \mathcal{N}_T \left(y(\cdot, t), \frac{\partial y(\cdot, t)}{\partial t} \right)^2$$

for all $t > 0$.

The following theorem says that a necessary and sufficient condition for \mathcal{N}_T to be a norm for some $T > 0$ is that the following holds:

(C1) The systems $(S_i), i = 1, \dots, N$, do not have a common eigenvalue.

In what follows, $T_i, i = 1, \dots, N$, will denote times for which the conclusion of Theorem 3.1 holds.

THEOREM 3.3. *If condition (C1) does not hold, then we can find nonzero $(f_1, f_2) \in H_0^1[0, L] \times L^2[0, L]$ such that $\mathcal{N}_T(f_1, f_2) = 0$ for all $T > 0$. Conversely, if condition (C1) holds, then $\mathcal{N}_T(\cdot, \cdot)$ defines a norm on $H_0^1[0, L] \times L^2[0, L]$ for any $T > T_0 = \max\{T_i + T_{i+1} : 1 \leq i \leq N-1\}$.*

Proof. Suppose that all the $(S_i), i = 1, \dots, N$ have a common eigenvalue λ and with eigenfunctions $\phi_i, i = 1, \dots, N$. Let $f_1(x) = c_i \phi_i(x)$ if $x \in (x_{i-1}, x_i)$ and let $f_2 \equiv 0$. With appropriate choice of the constants $c_i, i = 1, \dots, N$, we can make

$$c_i \tau_i(x_i) \phi_i'(x_i) = c_{i+1} \tau_{i+1}(x_i) \phi_{i+1}'(x_i), \quad i = 1, \dots, N-1.$$

Then

$$y(x, t) = c_i \cos(\lambda t) \phi_i(x) \quad \text{if } x \in (x_{i-1}, x_i)$$

would be the solution of (3.1)–(3.3) with initial condition $y(x, 0) = f_1(x), \partial y(x, 0)/\partial t = f_2(x)$. Since $y(x_i, t) = 0$, for $i = 1, \dots, N-1$ and $t > 0$, it follows that $\mathcal{N}_T(f_1, f_2) = 0$ for all $T > 0$.

Next, suppose that the system (S_i) has eigenvalues

$$\Lambda_i = \{\lambda_{n,i} : 1 \leq n < \infty\}$$

and normalized eigenfunctions (see (3.14))

$$\{\phi_{n,i}(x) : 1 \leq n < \infty\}$$

$i = 1, \dots, N$. Suppose $\mathcal{N}_T(f_1, f_2) = 0$. Let $y(x, t)$ be the solution of (3.1), (3.4). Then $\partial y(x_i, t)/\partial t = 0$ for $i = 1, \dots, N-1, 0 \leq t \leq T$. So we have $y(x_i, t) = \text{constant} = \alpha_i, i = 1, \dots, N-1$. Let

$$w(x, t) = y(x, t) - \alpha_{i-1} - \frac{(\alpha_i - \alpha_{i-1})(x - x_{i-1})}{x_i - x_{i-1}}.$$

(We take $\alpha_0 = \alpha_N = 0$.) Then w is a solution of $(S_i), i = 1, \dots, N$. So we can expand

$$(3.28) \quad w(x, t) = \sum_{n=1}^{\infty} (a_{n,i} e^{i\lambda_{n,i}t} + a_{-n,i} e^{i\lambda_{-n,i}t}) \phi_{n,i}(x),$$

where $\lambda_{-n,i} = -\lambda_{n,i}$. It follows that

$$\begin{aligned} & \sum_{n=1}^{\infty} (a_{n,i+1} e^{i\lambda_{n,i+1}t} + a_{-n,i+1} e^{i\lambda_{-n,i+1}t}) \phi'_{n,i+1}(x_i) \tau_{i+1}(x_1) \\ & - \sum_{n=1}^{\infty} (a_{n,i} e^{i\lambda_{n,i}t} + a_{-n,i} e^{i\lambda_{-n,i}t}) \phi'_{n,i}(x_i) \tau_i(x_i) \\ (3.29) \quad & = \tau_{i+1}(x_i) \frac{\partial}{\partial x} w(x_i^+, t) - \tau_i(x_i) \frac{\partial}{\partial x} w(x_i^-, t) \\ & = \tau_{i+1}(x_i) \frac{\partial}{\partial x} y(x_i^+, t) - \tau_i(x_i) \frac{\partial}{\partial x} y(x_i^-, t) - \tau_{i+1}(x_i) \frac{\alpha_{i+1} - \alpha_i}{x_{i+1} - x_i} \\ & \quad + \tau_i(x_i) \frac{\alpha_i - \alpha_{i-1}}{x_i - x_{i-1}} \\ & = -\tau_{i+1}(x_i) \frac{\alpha_{i+1} - \alpha_i}{x_{i+1} - x_i} + \tau_i(x_i) \frac{\alpha_i - \alpha_{i-1}}{x_i - x_{i-1}}. \end{aligned}$$

By Theorem 2.6, the set $\{e^{\pm i\lambda_{n,i}t} : 0 < n < \infty\} \cup \{e^{\pm i\lambda_{n,i+1}t} : 0 < n < \infty\} \cup \{1\}$ is strongly independent in $L^2[0, T]$. So by (3.29), we have

$$\begin{aligned} a_{n,i} &= 0 & \text{if } \lambda_{n,i} \notin \Lambda_i \cap \Lambda_{i+1}, \\ a_{n,i+1} &= 0 & \text{if } \lambda_{n,i+1} \notin \Lambda_i \cap \Lambda_{i+1}, \end{aligned}$$

and

$$(3.30) \quad \tau_{i+1}(x_i) \frac{\alpha_{i+1} - \alpha_i}{x_{i+1} - x_i} = \tau_i(x_i) \frac{\alpha_i - \alpha_{i-1}}{x_i - x_{i-1}}.$$

Suppose w is not identically zero. Then there exist some n and i such that $a_{n,i} \neq 0$. It follows that $\lambda_{n,i} \in \Lambda_{i+1}$. Let $\lambda_{n,i} = \lambda_{m,i+1}$. Then we must have $a_{m,i+1} \neq 0$. Hence $\lambda_{n,i} = \lambda_{m,i+1} \in \Lambda_{i+2}$. Proceeding in this way, we can then show that $\lambda_{n,i} \in \Lambda_j$ for all $j > i$. In a similar way, we can prove that $\lambda_{n,i} \in \Lambda_j$ for all $j < i$. Hence $\lambda_{n,i}$ is a common eigenvalue of the systems $(S_i), i = 1, \dots, N$, a contradiction. So we must have $w \equiv 0$. Also, since (3.30) holds for $i = 1, \dots, N-1$, it follows that $\alpha_i - \alpha_{i-1}, i = 1, \dots, N$, all have the same sign. Because $\alpha_0 = \alpha_N = 0$. This forces $\alpha_i = 0$ for $i = 1, \dots, N-1$. Hence we have proved that $y \equiv 0$ and therefore that $(f_1, f_2) = 0$. Thus \mathcal{N}_T is indeed a norm.

Remark. We see from equality (3.27) that if (C1) holds, then \mathcal{N}_T is a norm weaker than the $H_0^1[0, L] \times L^2[0, L]$ norm.

Theorem 3.4 is a consequence of Theorem 3.3.

THEOREM 3.4. *If condition (C1) does not hold, then the system (3.1)–(3.3) is not asymptotically stable. Conversely, if condition (C1) holds, then the solution of the system (3.1)–(3.3) satisfies*

$$\lim_{t \rightarrow +\infty} \mathcal{N} \left(y(\cdot, t), \frac{\partial y(\cdot, t)}{\partial t} \right) = 0,$$

where $\mathcal{N}(\cdot, \cdot)$ is a norm on $H_0^1[0, L] \times L^2[0, L]$.

Proof. The proof of the first part follows easily from the corresponding part in Theorem 3.3. Conversely, if the systems (S_i) do not have a common eigenvalue, then by Theorem 3.3, we can find some $T > 0$ such that \mathcal{N}_T is a norm on $H_0^1[0, L] \times L^2[0, L]$. From equality (3.24), we know that $\mathcal{E}y(t)$ is a nonnegative decreasing function of t . Hence $\alpha = \lim_{t \rightarrow +\infty} \mathcal{E}y(t)$ exists. Also, by (3.26), we have

$$\mathcal{N}_T \left(y(\cdot, t), \frac{\partial y(\cdot, t)}{\partial t} \right)^2 = \mathcal{E}y(t) - \mathcal{E}y(t + T).$$

Hence

$$\lim_{t \rightarrow +\infty} \mathcal{N}_T \left(y(\cdot, t), \frac{\partial y(\cdot, t)}{\partial t} \right)^2 = \alpha - \alpha = 0.$$

This completes the proof of the theorem.

We will use the following elementary result of functional analysis.

PROPOSITION 3.5. *If T is a bounded linear operator on a Hilbert space X satisfying*

$$\|\mathcal{F}x\| < \|x\| \quad \text{for all } x \in X, x \neq 0,$$

then $I - T$ has dense range.

THEOREM 3.6. *If condition (C1) does not hold, then the system (1.1)–(1.3) is not approximately controllable in any time $T > 0$. Conversely, if condition (C1) holds, then the system (1.1)–(1.3) is approximately controllable in any time $T > T_0 = \max\{t_i + T_{i+1} : 1 \leq i \leq N - 1\}$.*

Proof. Suppose that the systems (S_i) , $i = 1, \dots, N$, have a common eigenvalue λ with eigenfunctions ϕ_i , $i = 1, \dots, N$. Let

$$z(x, t) = c_i \phi_i(x) \cos \lambda t \quad \text{if } x \in [x_{i-1}, x_i], i = 1, \dots, N.$$

Then z satisfies

$$\begin{aligned} \frac{\partial^2 z}{\partial t^2} &= \frac{1}{\rho_i} \frac{\partial}{\partial x} \left(\tau_i \frac{\partial z}{\partial x} \right), & x \in (x_{i-1}, x_i); \\ z(0, t) &= z(L, t) = 0, \\ z(x_i^+, t) &= z(x_i^-, t) = 0, & i = 1, \dots, N - 1. \end{aligned}$$

Also, we can choose the constants c_i , $i = 1, \dots, N - 1$, so that

$$\tau_{i+1}(x_i) \frac{\partial z(x_i^+, t)}{\partial x} = \tau_i(x_i) \frac{\partial z(x_i^-, t)}{\partial x}, \quad i = 1, \dots, N - 1.$$

(See the proof of Theorem 3.3.) Let y be a solution of (1.1)–(1.3). It is easy to see that

$$\frac{d}{dt} \sum_{i=1}^{x_i} \rho_i \left(z \frac{\partial y}{\partial t} - y \frac{\partial z}{\partial t} \right) dx = 0.$$

So if y satisfies the initial condition

$$y(x, 0) = \frac{\partial y(x, 0)}{\partial t} = 0,$$

then

$$\sum_{i=1}^N \int_{x_{i-1}}^{x_i} \rho_i(x) \left(z(x, T) \frac{\partial y(x, T)}{\partial t} - y(x, T) \frac{\partial z(x, T)}{\partial t} \right) dx = 0.$$

Hence we have found a nonzero pair of functions orthogonal, in $L^2[0, L] \times L^2[0, L]$, to the set of terminal states $(y(x, T), \partial y(x, T)/\partial x)$. So this set is not dense in $L^2[0, L] \times L^2[0, L]$. Therefore, it is not dense in $H^1[0, L] \times L^2[0, L]$ either and the system is not approximately controllable in time T .

Conversely, suppose that the systems $(S_i), i = 1, \dots, N-1$ do not have a common eigenvalue. Then by Theorem 3.3, the seminorm \mathcal{N}_T is a norm if $T > T_0 = \max\{T_i + T_{i+1} : 1 \leq i \leq N-1\}$. Consider the norm $\|\cdot\|_E$ on $H_0^1[0, L] \times L^2[0, L]$ defined by

$$(3.31) \quad \|(f_1, f_2)\|_E = \left(\frac{1}{2} \sum_{k=1}^{N-1} \int_{x_i}^{x_{i+1}} \rho_i(x) f_1(x)^2 + \tau_i(x) f_2(x)^2 dx \right)^{1/2}.$$

This norm is equivalent with the usual $H_0^1[0, L] \times L^2[0, L]$ norm. Now consider given initial state (f_1, f_2) in $H_0^1[0, L] \times L^2[0, L]$. For (h_1, h_2) in $H_0^1[0, L] \times L^2[0, L]$, let y_1 be the solution of (3.1)–(3.3) with the initial conditions

$$y_1(x, 0) = h_1(x) \quad \text{and} \quad \frac{\partial y_1(x, 0)}{\partial t} = -h_2(x)$$

and let y_2 be the solution of (3.1)–(3.3) with initial conditions

$$y_2(x, 0) = y_1(x, T) - f_1(x) \quad \text{and} \quad \frac{\partial y_2(x, 0)}{\partial t} = -\frac{\partial y_1(x, T)}{\partial t} - f_2(x).$$

Let \mathcal{T} be the mapping that carries (h_1, h_2) to $(y_2(\cdot, T), \partial y_2(\cdot, T)/\partial t)$. It follows from equality (3.27) that because \mathcal{N}_T is a norm, we have $\|\mathcal{T}(h_1, h_2)\|_E < \|(h_1, h_2)\|_E$ for $(h_1, h_2) \neq 0$. Then by Proposition 3.5, $I - \mathcal{T}$ has dense range in $X = H_0^1[0, L] \times L^2[0, L]$. Now let $y(x, t) = y_1(x, T - t) - y_2(x, t)$. Then y is a solution of (1.1)–(1.3) with initial conditions

$$y(x, 0) = f_1(x) \quad \text{and} \quad \frac{\partial y(x, 0)}{\partial t} = f_2(x)$$

and control functions

$$u_i(t) = k_i \left(\frac{\partial y_1(x, T - t)}{\partial t} - \frac{\partial y_2(x, t)}{\partial t} \right), \quad i = 1, \dots, N-1.$$

Because both y_1 and y_2 have finite energy for all time t , by (3.25) we have $u_i \in L^2[0, T], i = 1, \dots, N-1$. Also, the terminal state of y is

$$y(x, T) = h_1(x) - y_2(x, T), \quad \frac{\partial y(x, T)}{\partial t} = h_2(x) - \frac{\partial y_2(x, T)}{\partial t}.$$

Because $I - \mathcal{T}$ has dense range, the set of such terminal states is dense in $H_0^1[0, L] \times L^2[0, L]$. Hence the system (1.1)–(1.3) is approximately controllable.

4. Stabilizability with a uniform exponential decay rate and exact controllability.

In this section, we still consider the closed-loop system (3.1)–(3.3) and the related systems of uncoupled strings $(S_i), i = 1, \dots, N$, with sets of eigenvalues $\{\lambda_{n,i} : 1 \leq n < \infty\}$ and eigenfunctions $\{\phi_{n,i} : 1 \leq n < \infty\}, i = 1, \dots, N$. We will give a sufficient condition under which the system (3.1)–(1.3) will decay with a uniform exponential rate. Then by the “controllability via stabilizability method” of Russell [15], this will, in turn, give us a sufficient condition for the open-loop system (1.1)–(1.3) to be exactly controllable. We consider the following condition:

(C2) There exists $i, 1 \leq i \leq N - 1$, and $T^* > 0$ such that

$$\{\lambda_{n,i} : 1 \leq n < \infty\} \cap \{\lambda_{n,i+1} : 1 \leq n < \infty\} = \emptyset$$

and the set of exponential functions

$$\{e^{\pm i\lambda_{n,i}t} : 1 \leq n < \infty\} \cup \{e^{\pm i\lambda_{n,i+1}t} : 1 \leq n < \infty\}$$

satisfies condition (RB2) in $L^2[0, T^*]$.

Clearly (C2) is a stronger condition than (C1). Let \mathcal{N}_T and T_1, \dots, T_N be as defined in the previous section.

THEOREM 4.1. *If condition (C2) holds, then the seminorm \mathcal{N}_T is a norm equivalent with the $H_0^1[0, L] \times L^2[0, L]$ norm if $T > T_0 = \max\{T^*, T_1, T_2, \dots, T_N\}$.*

To prove Theorem 4.1, we need the following lemma.

LEMMA 4.2. *Let $T > 0$. There exists a positive constant K such that for any v satisfying*

$$(4.1) \quad \rho \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left(\tau \frac{\partial v}{\partial x} \right), \quad t > 0, a < x < b;$$

$$(4.2) \quad v(x, 0) = \frac{\partial v(x, 0)}{\partial t} = 0, \quad a \leq x \leq b;$$

we have

$$(4.3) \quad \int_0^T \left(\frac{\partial v(a, t)}{\partial x} \right)^2 + \left(\frac{\partial v(b, t)}{\partial x} \right)^2 dt \leq K \int_0^T \left(\frac{\partial v(a, t)}{\partial t} \right)^2 + \left(\frac{\partial v(b, t)}{\partial t} \right)^2 dt.$$

Proof. By a multiplier method, (using, for example, the multiplier $(x - (a+b)/2)\partial w/\partial x$) we can easily prove the following inequality:

$$(4.4) \quad \int_0^s \left(\frac{\partial w(a, t)}{\partial x} \right)^2 + \left(\frac{\partial w(a, t)}{\partial t} \right)^2 + \left(\frac{\partial w(b, t)}{\partial x} \right)^2 + \left(\frac{\partial w(b, t)}{\partial t} \right)^2 dt \\ \leq K_1 \left(\mathcal{E}w(0) + \mathcal{E}w(s) + \int_0^s \mathcal{E}w(t) dt \right)$$

for any function w satisfying

$$(4.5) \quad \rho \frac{\partial^2 w}{\partial t^2} = \frac{\partial}{\partial x} \left(\tau \frac{\partial w}{\partial x} \right), \quad 0 \leq t \leq s, \quad a \leq x \leq b.$$

Here K_1 is a constant independent of w and s . Suppose that in addition,

$$(4.6) \quad w(a, t) = k_1, \quad w(b, t) = k_2,$$

for some constants k_1 and k_2 . Then by conservation of energy and (4.4) we have

$$(4.7) \quad \int_0^s \left(\frac{\partial w(a, t)}{\partial x} \right)^2 + \left(\frac{\partial w(b, t)}{\partial x} \right)^2 dt \leq K_1(2+s)\mathcal{E}w(0) = K_1(2+s)\mathcal{E}w(s).$$

If v satisfies (4.1), (4.2), then

$$\begin{aligned} & \frac{d}{dt} \int_a^b \rho \frac{\partial w}{\partial t} \frac{\partial v}{\partial t} + \tau \frac{\partial w}{\partial x} \frac{\partial v}{\partial x} dx \\ &= \tau(b) \frac{\partial w(b, t)}{\partial x} \frac{\partial v(b, t)}{\partial t} - \tau(a) \frac{\partial w(a, t)}{\partial x} \frac{\partial v(a, t)}{\partial t}. \end{aligned}$$

Hence

$$\begin{aligned} & \int_a^b \rho(x) \frac{\partial w(x, s)}{\partial t} \frac{\partial v(x, s)}{\partial t} + \sigma(x) \frac{\partial w(x, s)}{\partial x} \frac{\partial v(x, s)}{\partial x} dx \\ &= \tau(b) \int_0^s \frac{\partial w(b, t)}{\partial x} \frac{\partial v(b, t)}{\partial t} dt \\ &\quad - \tau(a) \int_0^s \frac{\partial w(a, t)}{\partial x} \frac{\partial v(a, t)}{\partial t} dt \\ &\leq \tau(a) \left(\int_0^s \left(\frac{\partial w(a, t)}{\partial x} \right)^2 dt \right)^{1/2} \left(\int_0^s \left(\frac{\partial v(a, t)}{\partial t} \right)^2 dt \right)^{1/2} \\ &\quad + \tau(b) \left(\int_0^s \left(\frac{\partial w(b, t)}{\partial x} \right)^2 dt \right)^{1/2} \left(\int_0^s \left(\frac{\partial v(b, t)}{\partial t} \right)^2 dt \right)^{1/2} \\ &\leq m(K_1(2+s)\mathcal{E}w(s))^{1/2} \left(\int_0^s \left(\frac{\partial v(a, t)}{\partial t} \right)^2 \right. \\ &\quad \left. + \left(\frac{\partial v(b, t)}{\partial t} \right)^2 dt \right)^{1/2}, \end{aligned}$$

where $m = \max\{\tau(a), \tau(b)\}$. Since this holds for any w satisfying (4.5) and (4.6), we conclude that

$$(4.8) \quad \mathcal{E}v(s) \leq K_2(2+s) \int_0^s \left(\frac{\partial v(a, t)}{\partial t} \right)^2 + \left(\frac{\partial v(b, t)}{\partial t} \right)^2 dt,$$

where K_2 is a constant independent of v and s . Noting that this holds for any arbitrary s , we have, by (4.4) again, that

$$\begin{aligned} & \int_0^T \left(\frac{\partial v(a, t)}{\partial x} \right)^2 + \left(\frac{\partial v(b, t)}{\partial x} \right)^2 dt \\ &\leq K_1 \left(\int_0^T \mathcal{E}v(t) dt + \mathcal{E}v(T) \right) \\ &\leq K_1 K_2 \left(\int_0^T (2+t) \int_0^t \left(\frac{\partial v(a, s)}{\partial t} \right)^2 + \left(\frac{\partial v(b, s)}{\partial t} \right)^2 ds dt \right. \\ &\quad \left. + (2+T) \int_0^T \left(\frac{\partial v(a, t)}{\partial t} \right)^2 + \left(\frac{\partial v(b, t)}{\partial t} \right)^2 dt \right) \end{aligned}$$

$$\leq K_1 K_2 (1+T)(2+T) \int_0^T \left(\frac{\partial v(a,t)}{\partial t} \right)^2 + \left(\frac{\partial v(b,t)}{\partial t} \right)^2 dt.$$

This completes the proof of the lemma.

Proof of Theorem 4.1. Let y satisfy (3.1)–(3.3) with initial condition (3.4). Let v satisfy

$$\begin{aligned} \rho_i \frac{\partial^2 v}{\partial t^2} &= \frac{\partial}{\partial x} \left(\tau_i \frac{\partial v}{\partial x} \right), \quad t > 0, \quad x_{i-1} < x < x_i, \quad i = 1, \dots, N; \\ v(x, 0) &= \frac{\partial v(x, 0)}{\partial t} = 0, \quad \text{for } 0 < x < L; \\ v(x_i^-, t) &= v(x_i^+, t) = y(x, t), \quad i = 1, \dots, N-1; \\ v(0, t) &= v(L, t) = 0. \end{aligned}$$

Then $w = y - v$ satisfies

$$\begin{aligned} \rho_i \frac{\partial^2 w}{\partial t^2} &= \frac{\partial}{\partial x} \left(\tau_i \frac{\partial w}{\partial x} \right), \quad t > 0, \quad x_{i-1} < x < x_i, \quad i = 1, \dots, N; \\ w(x, 0) &= f_1(x), \quad \frac{\partial w(x, 0)}{\partial t} = f_2(x) \quad \text{for } 0 < x < L; \\ w(x_i^-, t) &= w(x_i^+, t) = 0, \quad i = 1, \dots, N-1; \\ w(0, t) &= w(L, t) = 0. \end{aligned}$$

Thus w is a solution of the uncoupled strings $(S_i), i = 1, \dots, N$. Let $1 \leq j \leq N-1$ be such that for $i = j$, condition (C2) is satisfied. We expand

$$(4.9) \quad w(x, t) = \sum_{n=1}^{\infty} (a_{n,i} e^{i\lambda_{n,i}t} + a_{-n,i} e^{i\lambda_{-n,i}t}) \phi_{n,i}(x).$$

Denote $\omega_i(t) = \tau_{i+1}(x_i)(\partial/\partial x)w(x_i^+, t) - \tau_i(x_i)(\partial/\partial x)w(x_i^+, t)$. Then

$$\begin{aligned} \omega_i(t) &= \sum_{n=1}^{\infty} (a_{n,i+1} e^{i\lambda_{n,i+1}t} + a_{-n,i+1} e^{i\lambda_{-n,i+1}t}) \phi'_{n,i+1}(x_i) \tau_{i+1}(x_i) \\ &\quad - \sum_{n=1}^{\infty} (a_{n,i} e^{i\lambda_{n,i}t} + a_{-n,i} e^{i\lambda_{-n,i}t}) \phi'_{n,i}(x_i) \tau_i(x_i). \end{aligned}$$

By the observability result Theorem 3.1 and the fact that the eigenfunctions are normalized, it follows that for all $i, 1 \leq i \leq N$, the sequences $\{\phi'_{n,i}(x_{i-1})\}_{n=1}^{\infty}$ and $\{\phi'_{n,i}(x_i)\}_{n=1}^{\infty}$ are bounded and bounded away from zero. So since the set of exponential functions

$$\{e^{\pm i\lambda_{n,j}t} : 1 \leq n < \infty\} \cup \{e^{\pm i\lambda_{n,j+1}t} : 1 \leq n < \infty\}$$

satisfies condition (RB2) in $L^2[0, T]$, there exist constant $d > 0$ such that

$$(4.10) \quad \int_0^{T^*} |\omega_j(t)|^2 dt \geq d \sum_{n=1}^{\infty} |a_{n,j+1}|^2 + |a_{-n,j+1}|^2 + |a_{n,j}|^2 + |a_{-n,j}|^2.$$

Hence, there exists constant $d_1 > 0$ such that

$$(4.11) \quad \int_0^{T^*} |\omega_j(t)|^2 dt \geq d_1 (E_{j+1} + E_j),$$

where E_i denotes the total (constant) energy of the system $(S_i), i = 1, \dots, N$. By the observability result Theorem 3.1, it follows that for each $i, 1 \leq i \leq N$, there exists constant $K_i > 0$ such that

$$E_i \leq K_i \int_0^{T_i} \left| \frac{\partial}{\partial x} w(x_{i-1}^+, t) \right|^2 dt$$

and

$$E_i \leq K_i \int_0^{T_i} \left| \frac{\partial}{\partial x} w(x_i^-, t) \right|^2 dt.$$

and, by regularity results [10], there exists constant $M_i > 0$ such that

$$\int_0^{T_i} \left| \frac{\partial}{\partial x} w(x_{i-1}^+, t) \right|^2 dt \leq M_i E_i$$

and

$$\int_0^{T_i} \left| \frac{\partial}{\partial x} w(x_i^-, t) \right|^2 dt \leq M_i E_i$$

So for $i, j+1 < i \leq N$,

$$\begin{aligned} E_i &\leq K_i \int_0^{T_i} \left| \frac{\partial}{\partial x} w(x_{i-1}^+, t) \right|^2 dt \\ (4.12) \quad &\leq 2K_i/\tau_i(x_{i-1})^2 \left(\int_0^{T_i} \left| \tau_{i-1}(x_{i-1}) \frac{\partial}{\partial x} w(x_{i-1}^-, t) \right|^2 + |\omega_{i-1}(t)|^2 dt \right) \\ &\leq 2K_i/\tau_i(x_{i-1})^2 \left(M_i \tau_{i-1}(x_{i-1})^2 E_{i-1} + \int_0^{T_i} |\omega_{i-1}(t)|^2 dt \right). \end{aligned}$$

Hence, we have

$$(4.13) \quad E_{j+2} + \dots + E_N \leq C_1 \left(\sum_{i=j+1}^{N-1} \int_0^{T_i} |\omega_i(t)|^2 dt + E_{j+1} \right)$$

for some positive constant C_1 . In a similar way, we can prove that

$$(4.14) \quad E_1 + \dots + E_{j-1} \leq C_2 \left(\sum_{i=1}^{j-1} \int_0^{T_i} |\omega_i(t)|^2 dt + E_j \right)$$

for some positive constant C_2 . Combining (4.11), (4.13), and (4.14), we have

$$(4.15) \quad \|(f_1, f_2)\|_E^2 = \mathcal{E}w(0) = E_1 + E_2 + \dots + E_N \leq C_3 \sum_{i=1}^N \int_0^{T_i} |\omega_i(t)|^2 dt$$

for some positive constant C_3 . But by the definition of ω_i and the conditions on y at the coupled points, we have

$$\omega_i(t) = k_i \frac{\partial}{\partial t} y(x_i, t) - \left(\tau_{i+1}(x_i) \frac{\partial}{\partial x} v(x_i^+, t) - \tau_i(x_i) \frac{\partial}{\partial x} v(x_i^-, t) \right).$$

Hence

$$(4.16) \quad \sum_{i=1}^N \int_0^{T_i} |\omega_i(t)|^2 dt \leq C_4 \sum_{i=1}^N \int_0^{T_i} \left| \frac{\partial y(x_i, t)}{\partial t} \right|^2 + \left| \frac{\partial v(x_{i-1}^+, t)}{\partial x} \right|^2 + \left| \frac{\partial v(x_i^-, t)}{\partial x} \right|^2 dt$$

for some constant $C_4 > 0$. By Lemma 4.2 and the fact that $\partial v(x_i, t)/\partial t = \partial y(x_i, t)/\partial t$, we have

$$(4.17) \quad \sum_{i=1}^N \int_0^{T_i} |\omega_i(t)|^2 dt \leq C_5 \sum_{i=1}^N \int_0^{T_i} \left| \frac{\partial y(x_i, t)}{\partial t} \right|^2 dt \leq C_6 \mathcal{N}_T(f_1, c_2),$$

where C_5 and C_6 are positive constants. Combining (4.15) and (4.17), we see that the norm \mathcal{N}_T is stronger than the $H_0^1[0, L] \times L^2[0, L]$ norm. Since we already know that the norm \mathcal{N}_T is weaker than the $H_0^1[0, L] \times L^2[0, L]$ norm (see the remark after Theorem 3.3) this completes the proof of Theorem 4.1.

From Theorem 4.1, we immediately have the following theorem.

THEOREM 4.3. *If condition (2) holds, then the system (3.1)–(3.3) is asymptotically stable with a uniform exponential rate of decay. In other words, we can find positive constants M and k such that*

$$\mathcal{E}y(t) \leq M e^{-kt} \mathcal{E}y(0)$$

for all y satisfying (3.1)–(3.3).

Proof. Suppose (C2) holds. Let T_0 be as in the statement of Theorem 4.1. Then for $T > T_0$, \mathcal{N}_T is equivalent with the $H_0^1[0, L] \times L^2[0, L]$ norm. Hence we can find a constant $c > 0$ such that

$$(4.18) \quad \mathcal{N}_T(f_1, f_2) \geq c \|(f_1, f_2)\|_E.$$

Then by (3.27), we have, for any $t > 0$,

$$(4.19) \quad \begin{aligned} \mathcal{E}y(t+T) - \mathcal{E}y(t) &= -\mathcal{N}_T \left(y(\cdot, t), \frac{\partial y(\cdot, t)}{\partial t} \right)^2 \\ &\leq -c^2 \left\| \left(y(\cdot, t), \frac{\partial y(\cdot, t)}{\partial t} \right) \right\|_E^2 = -c^2 \mathcal{E}y(t). \end{aligned}$$

Because $\mathcal{E}y(t)$ decreases with t , it follows from the semigroup property that

$$\mathcal{E}y(t) \leq M e^{-kt} \mathcal{E}y(0),$$

where $M = (1 - c^2)^{-1}$ and $k = -\ln M/T$. (Note that from (4.19), we must have $c^2 < 1$.)

Using the “controllability via stabilizability method” of Russell, we can then prove the following theorem.

THEOREM 4.4. *If condition (C2) holds, then the system is exactly controllable in any time $T > T_0 \max\{T^*, T_1, T_2, \dots, T_n\}$.*

5. Examples. We consider a special situation where $N = 2$ and ρ_i, τ_i are constants, $i = 1, 2$. Denoting $c_i = (\tau_i/\rho_i)^{1/2}$ and $L_i = x_i - x_{i-1}$, $i = 1, 2$, the system (1.1)–(1.3) becomes

$$(5.1) \quad \frac{\partial^2 y}{\partial t^2} = c_i^2 \frac{\partial^2 y}{\partial x^2}, \quad x \in (x_{i-1}, x_i), \quad i = 1, 2;$$

$$(5.2) \quad y(0, t) = y(L, t) = 0,$$

$$(5.3) \quad y(x_1^+, t) = y(x_1^-, t),$$

$$(5.4) \quad \tau_2 \frac{\partial y(x_1^+, t)}{\partial x} - \tau_1 \frac{\partial y(x_1^-, t)}{\partial x} = u_1(t).$$

The eigenvalues of the uncoupled strings are

$$\lambda_{n,i} = n\pi c_i/L_i, \quad n = 1, 2, 3, \dots, \quad \text{and} \quad i = 1, 2.$$

It follows that $\lambda_{n,1} = \lambda_{m,2}$ for some m, n if and only if $c_1 L_2/c_2 L_1$ is rational. Hence from Theorem 3.6, we conclude that *the system (5.1)–(5.4) is approximately controllable (in some time $T > 0$) if and only if $c_1 L_2/c_2 L_1$ is irrational*. The time T can be any number greater than

$$(5.5) \quad T_0 = 2 \left(\frac{L_1}{c_1} + \frac{L_2}{c_2} \right).$$

However, we can show that even when $c_1 L_2/c_2 L_1$ is irrational, *the system is not exactly controllable*.

To prove that, let us first consider a function defined by the formula

$$f_\mu(x) = \begin{cases} \sin \frac{\mu L_2}{c_2} \sin \frac{\mu x}{c_1} & \text{if } 0 < x < x_1, \\ \sin \frac{\mu L_1}{c_1} \sin \frac{\mu(L-x)}{c_2} & \text{if } x_1 < x < L. \end{cases}$$

Here μ is a real parameter. It is easy to verify that

$$\tilde{y}(x, t) = \sin \mu(T-t) f_\mu(x)$$

is a solution of (5.1)–(5.4) with

$$\tilde{u}_1(t) = -\mu \frac{F(\mu) \sin \mu(T-t)}{c_1 c_2},$$

where

$$F(\mu) = c_1 \tau_1 \sin \frac{\mu L_1}{c_2} \cos \frac{\mu L_2}{c_2} + c_2 \tau_2 \sin \frac{\mu L_2}{c_2} \cos \frac{\mu L_1}{c_1}.$$

If y is any solution of (5.1)–(5.4), we have

$$(5.6) \quad \frac{d}{dt} \sum_{i=1}^2 \int_{x_{i-1}}^{x_i} \rho_i \frac{\partial y}{\partial t} \frac{\partial \tilde{y}}{\partial t} + \tau_i \frac{\partial y}{\partial x} \frac{\partial \tilde{y}}{\partial x} dx = -\frac{\partial y(x_1, t)}{\partial t} \tilde{u}_1(t) - \frac{\partial \tilde{y}(x_1, t)}{\partial t} u_1(t).$$

Note that if $F(\mu) = 0$, then $\tilde{u}_1(t) = 0$. We now construct a sequence $\{\mu_n\}_{n=0}^\infty$ such that $F(\mu_n) = 0$ for all n .

From a result in number theory, we know that if $c_1 L_2 / c_2 L_1$ is an irrational number, there exist two sequences of integers $\{p_n\}_{n=0}^{\infty}$ and $\{q_n\}_{n=0}^{\infty}$ tending to infinity as n tends to infinity, such that

$$(5.7) \quad \left| \frac{c_1 L_2}{c_2 L_1} - \frac{p_n}{q_n} \right| < 1/q_n^2$$

for all n . (See, e.g., Hardy and Wright [3, Thm. 171, p. 140].) Let $\lambda_n = p_n c_2 \pi / L_2$. Then

$$(5.8) \quad \sin \frac{\lambda_n L_2}{c_2} = 0.$$

also, by (5.7), we have

$$\left| \frac{\lambda_n L_1}{c_1} - q_n \pi \right| < \epsilon_n,$$

where

$$\epsilon_n = \frac{c_2 L_1 \pi}{c_1 L_2 q_n}.$$

So

$$(5.9) \quad \left| \sin \frac{\lambda_n L_1}{c_1} \right| < \epsilon_n \quad \text{for all } n.$$

It follows from (5.8) and (5.9) that we have

$$(5.10) \quad |F(\lambda_n)| < c_1 \tau_1 \epsilon_n \quad \text{for all } n$$

and

$$(5.11) \quad \lim_{n \rightarrow \infty} |F'(\lambda_n)| = \tau_1 L_1 + \tau_2 L_2.$$

Because F'' is a bounded function, we can then find a constant $\delta > 0$, independent of n , such that for all n sufficiently large,

$$(5.12) \quad |F'(\lambda)| > (\tau_1 L_1 + \tau_2 L_2)/2 \quad \text{whenever } |\lambda - \lambda_n| < \delta.$$

From (5.10), (5.12) and the fact that $\lim_{n \rightarrow \infty} \epsilon_n = 0$, we conclude that for all n sufficiently large, there exists μ_n such that

$$(5.13) \quad |\mu_n - \lambda_n| < \frac{2c_1 \tau_1 \epsilon_n}{\tau_1 L_1 + \tau_2 L_2}$$

and

$$(5.14) \quad F(\mu_n) = 0.$$

Let

$$g_n(x) = f_{\mu_n}(x).$$

Let y be a solution of (5.1)–(5.4) with zero initial conditions. Then setting $\mu = \mu_n$ (which makes $\tilde{u}_1 = 0$) and integrating (5.6) from $t = 0$ to $t = T$ and canceling μ_n , we have

$$(5.15) \quad -\sum_{i=1}^2 \int_{x_{i-1}}^{x_i} \rho_i \frac{\partial y(x, T)}{\partial t} g_n(x) dx = g_n(x_1) \int_0^T \cos \mu_n(T-t) u_1(t) dt.$$

If the system (5.1)–(5.4) is exactly controllable, there exists a bounded linear mapping \mathcal{C} from $H_0^1[0, L] \times L^2[0, L]$ into $L^2[0, T]$ that maps the terminal state $(y(\cdot, T), \partial y(\cdot, T)/\partial t)$ to a control function u_1 that steers the zero initial state to this terminal state. Hence (5.15) implies that

$$\begin{aligned} \sum_{i=1}^2 \int_{x_{i-1}}^{x_i} \rho_i \frac{\partial y(x, T)}{\partial t} g_n(x) dx &\leq K |g_n(x_1)| \left(\mathcal{E}y(T) \int_0^T |\cos \mu_n(T-t)|^2 dt \right)^{1/2} \\ &= K |g_n(x_1)| \left(\mathcal{E}y(T) \left(\frac{T}{2} - \frac{\sin 2\mu_n(T-t)}{4\mu_n} \right) \right)^{1/2} \\ &\leq K |g_n(x_1)| \left(\mathcal{E}y(T) \left(\frac{T}{2} + \frac{1}{4\mu_n} \right) \right)^{1/2}. \end{aligned}$$

Here K is a positive constant. Since the terminal state can be arbitrary, this implies that there exists a constant K_1 such that

$$\sum_{i=1}^2 \int_{x_{i-1}}^{x_i} |g_n(x)|^2 dx \leq K_1 |g_n(x_1)|^2.$$

In other words

$$\begin{aligned} \left(\sin \frac{\mu_n L_2}{c_2} \right) \left(\frac{L_1}{2} - \frac{\sin(2\mu_n L_1/c_1)}{4\mu_n/c_1} \right) + \left(\sin \frac{\mu_n L_1}{c_1} \right)^2 \left(\frac{L_2}{2} - \frac{\sin(2\mu_n L_2/c_2)}{4\mu_n/c_2} \right) \\ \leq K_1 \left(\sin \frac{\mu_n L_1}{c_1} \right)^2 \left(\sin \frac{\mu_n L_2}{c_2} \right). \end{aligned}$$

(5.16)

For n sufficiently large, both terms on the left-hand side of the above inequality are nonnegative. Furthermore, because $c_1 L_2 / c_2 L_1$ is irrational, $\sin(\mu_n L_1 / c_1)$ and $\sin(\mu_n L_2 / c_2)$ cannot vanish simultaneously. Suppose there exist infinitely many n such that $\sin(\mu_n L_2 / c_2) \neq 0$. For such n , n sufficiently large, we must have

$$(5.17) \quad \frac{L_1}{2} - \frac{\sin(2\mu_n L_1/c_1)}{4\mu_n/c_1} \leq K_1 \left(\sin \frac{\mu_n L_1}{c_1} \right)^2.$$

But the inequalities (5.9) and (5.13) together imply that

$$\lim_{n \rightarrow \infty} \sin \frac{\mu_n L_1}{c_1} = 0.$$

So letting n tend to infinity in (5.17), we have $L_1/2 \leq 0$, a contradiction. In a similar way, we can show that if there exist infinitely many n such that $\sin(\mu_n L_2 / c_2) \neq 0$, then $L_2/2 \leq 0$, again a contradiction. So the system (5.1)–(5.4) cannot be exactly controllable.

The situation would be different if we change the boundary condition at one of the endpoints (say $x = L$) to

$$(5.18) \quad \frac{\partial}{\partial x} z(L, t) = 0$$

but keeping the boundary condition at the other end the same as before. With obvious modifications in their proofs, we see that Theorems 3.6 and 4.4 remain valid when the boundary condition at $x = L$ is replaced by the condition (5.18). Now, the eigenvalues of the corresponding uncoupled strings are

$$\lambda_{n,1} = n\pi c_1/K_1 \quad \text{and} \quad \lambda_{n,2} = (n - \tfrac{1}{2})\pi c_2/L_2, \quad n = 1, 2, \dots$$

Obviously, if $c_1 L_2/c_2 L_1$ is irrational, $\lambda_{n,1} \neq \lambda_{m,2}$ for any n and m . So the system (5.1)–(5.4) is approximately controllable in time T , with $T > T_0$, T_0 again given by (5.5). But note that if

$$(5.19) \quad \frac{c_1 L_2}{c_2 L_1} = \text{an integer/an odd integer} = \frac{r}{2s-1},$$

where r and s are positive integers, then

$$|\lambda_{n,1} - \lambda_{m,2}| = \left| \frac{\pi c_2}{L_2 2(2s-1)} (2nr - (2m-1)(2s-1)) \right| \geq \frac{\pi c_2}{L_2 2(2s-1)} > 0$$

for any n and m . Therefore, the system (5.1)–(5.4) is approximately controllable in time $T > T_0$, T_0 given by (5.5). Furthermore, all the $\lambda_{n,1}$ and $\lambda_{n,2}$, $n = 1, 2, \dots$, are multiples of a fixed number

$$\alpha = (\pi c_2)/(2L_2(2s-1)).$$

So the set of exponential functions

$$S = \{e^{\pm i\lambda_{n,1}t} : 1 \leq n < \infty\} \cup \{e^{\pm i\lambda_{n,2}t} : 1 \leq n < \infty\}$$

is a subset of the exponential functions

$$\{e^{in\alpha t} : -\infty < n < \infty\},$$

which is orthogonal in $L^2[0, 2\pi/\alpha]$. Hence so is S . So S satisfies (RB2). It follows that if assumption (5.19) is satisfied, then the system (5.1)–(5.4) is exactly controllable in any time T such that

$$T > T^* = \max \left\{ 2 \left(\frac{L_1}{c_1} + \frac{L_2}{c_2} \right), \frac{2\pi}{\alpha} \right\} = \max \left\{ 2 \left(\frac{L_1}{c_1} + \frac{L_2}{c_2} \right), \frac{4L_2(2s-1)}{c_2} \right\}.$$

Because $L_2(2s-1)/c_2 = L_1 r/c_1$, it follows that

$$\frac{L_2(2s-1)}{c_2} = \frac{1}{2} \left(\frac{L_2(2s-1)}{c_2} + \frac{L_1 r}{c_1} \right) \geq \frac{1}{2} \left(\frac{L_1}{c_1} + \frac{L_2}{c_2} \right).$$

Hence $T^* = 4(L_2(2s-1)/c_2)$. If either one of s or r is greater than 1, then T^* is greater than T_0 , the infimum time for approximate controllability. Of course, we have not shown that the time T^* given in Theorem 4.4 is optimal. So we do not know whether we actually

need a greater time for exact controllability than for just approximate controllability, when r or s is greater than 1.

The results in the above examples are consistent with those in [2] and [11]. For example, it was shown in [11] that for the case of “symmetric” boundary conditions

$$(5.20) \quad z(0, t) = z(2, t) = 0,$$

with feedback

$$(5.21) \quad z_t(1^+, t) - z_t(1^-, t) = -K_1 \tau_1 u_x(1^-, t), \quad K_1 > 0$$

the system is asymptotically stable if and only if the ratio of the wave speeds is irrational. (In [11], Liu took $L_1 = L_2 = 1, L = 2$.) However, the system never decays with a uniform exponential rate.

For the case of “unsymmetric” boundary conditions

$$(5.22) \quad z(0, t) - \frac{\partial}{\partial x} z(2, t) = 0$$

the system is asymptotically stable, with feedback (5.21), if and only if

$$c_1/c_2 = \text{an odd integer/an even integer.}$$

However, it decays with a uniform exponential rate if

$$c_1/c_2 = \text{an integer/an odd integer.}$$

This last condition is, of course, the same as (5.19).

REFERENCES

- [1] C. BARDOS, G. LEBEAU, AND J. RAUCH, *Contrôle et stabilisation dans les problèmes hyperboliques*, appendix in *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Tome 1, J. L. Lions, Masson, Paris, 1988.
- [2] G. CHEN, M. COLEMAN, AND H. H. WEST, *Pointwise stabilizability in the middle of the span for second order systems, nonuniform and uniform exponential decay of solutions*, SIAM J. Appl. Math., 47 (1987), pp. 751–780.
- [3] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, 4th ed., Oxford University Press, London, 1960.
- [4] L. F. HO, *Observabilité frontière de l'équation des ondes*, C. R. Acad. Sci. Paris, 302 (1986), pp. 443–446.
- [5] ———, *Exact controllability of second order hyperbolic systems with control in the Dirichlet boundary conditions*, J. Math. Pures et Appl., 66 (1987), pp. 363–368.
- [6] ———, *Exact controllability of the one-dimensional wave equation with locally distributed control*, SIAM J. Control Optim., 28 (1990), pp. 733–748.
- [7] V. KORMONIK, *Contrôlabilité exacte en temps minimal*, C. R. Acad. Sci. Paris, 304 (1987), pp. 605–608.
- [8] N. LEVINSON, *Gap and Density Theorems*, Amer. Math. Soc. Colloq. Publ., Vol. 26, Amer. Math. Soc., Providence, 1940.
- [9] N. LEVINSON AND C. MCCALLA, *Completeness and Independence of the solutions of some functional differential equations*, Stud. Appl. Math., 53 (1974) pp. 1–15.
- [10] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Non-homogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures et Appl., 65 (1986), pp. 149–192.
- [11] K. S. LIU, *Energy decay problem in the design of a point stabilizer for string vibrating systems*, SIAM J. Control Optim., 26 (1988), pp. 1348–1256.
- [12] K. S. LIU, F. L. HUANG, AND G. CHEN, *Exponential stability analysis of a long chain of coupled vibrating strings with dissipative linkage*, SIAM J. Appl. Math., 49 (1989), pp. 1694–1707.
- [13] F. RIESZ AND B. SZ. NAGY, *Functional Analysis*, Ungar, New York, 1955.
- [14] D. L. RUSSELL, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.
- [15] ———, *Controllability and stabilization theory for linear partial differential equations. Recent progress and open questions*, SIAM Review, 20 (1978), pp. 639–739.

SIMULTANEOUS COEFFICIENT ASSIGNMENT OF DISCRETE-TIME MULTI-INPUT MULTI-OUTPUT LINEAR TIME-VARYING SYSTEM: A NEW APPROACH FOR COMPENSATOR DESIGN*

BIJOY K. GHOSH[†] AND PAUL R. BOUTHELLIER[‡]

Abstract. In this paper, a linear time-varying input-output system is considered and its realization as a linear time-varying autoregressive moving average system (ARMA) is studied. A time-varying z -transform is also introduced and its properties are studied. Furthermore a time-varying version of the coefficient assignment problem well known in time invariant system theory as the pole placement problem is posed and analyzed. A r -tuple of discrete time, linear time-varying plants with m inputs and p outputs are considered together with a single p input m output linear time-varying compensator. The design objective is to construct a single compensator that “coefficient assigns,” and hence “bounded input bounded output stabilizes” under suitable additional technical assumptions, the set of r plants simultaneously in the closed loop. Such a problem is useful in robust design of linear time-varying control systems in the closed loop. Among the results, it is shown that a generic r -tuple of $p \times m$ plants (in a suitable topology) is simultaneously coefficient assignable, provided that $r < m/p$. The design procedure involves splitting the closed-loop system into an ARMA system in cascade with a moving average system. The coefficient assignment problem consists of assigning the coefficients of the autoregressive part of the ARMA subsystem. Thereby an algorithm is obtained that is nonrecursive and involves solving for each time instant a system of linear equations with time-varying coefficients. The associated time-varying matrix has the “Sylvester matrix structure.” Such a structure is well-known in pole placement of time-invariant systems by dynamic compensation. Additionally the problem of coefficient assignment of the autoregressive part of the ARMA system is considered in the closed loop, without splitting up into a cascade of two subsystems as before. A new recursive algorithm to analyze this problem has been introduced. The proposed algorithm has no counterpart in the time-invariant system design and thus represents a new design procedure. A special case of this algorithm for the single-input single-output system has been described in detail. An interesting feature of the proposed recursive algorithm is that one obtains a nonlinear recursion on the compensator parameters that would assign a prespecified sequence of coefficients for the closed-loop system. For a specific design problem it is shown that the dynamics of this nonlinear recursion is chaotic.

Key words. coefficient assignment, recursive algorithm, simultaneous design, time-varying system, chaotic dynamics

AMS(MOS) subject classifications. 14, 93

1. Introduction. Motivated by earlier successes in the field of simultaneous system design (see [1]–[7]) for linear time-invariant systems, in this paper the same general idea is applied to linear time-varying (LTV) systems as well. The motivation for considering time-varying systems is as follows. Many systems are time varying because they switch modes frequently (namely, high-performance aircrafts, power systems undergoing several modes of failure, etc.). Time-varying systems also arise from nonlinear systems linearized along a nominal trajectory. Furthermore, time-varying systems also arise from linear systems, where the parameters are perturbed by a time-varying function. A feedback design strategy that leads to a time-varying system is *adaptive control*, wherein the time variation is a result of real time adaptation. An important pair of problems in the design of time-varying systems is described as follows.

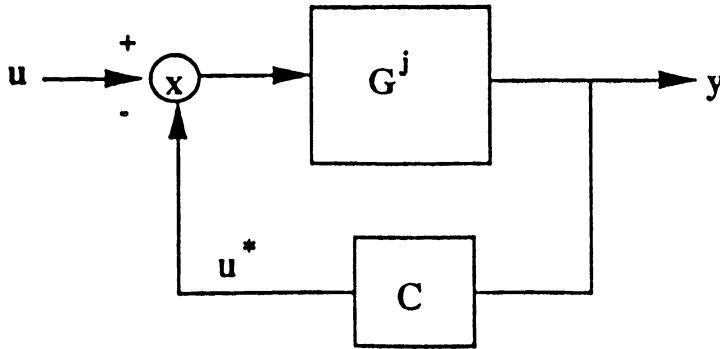
PROBLEM 1.1 (stability criterion). Given a class of linear time-varying systems. What condition on the parameters of the systems would guarantee bounded-input bounded-output (BIBO) stability?

PROBLEM 1.2 (stability criterion). If a linear time-varying system is not already BIBO stable, find a linear time-varying output feedback compensator such that the closed-loop system is BIBO stable.

* Received by the editors September 12, 1990; accepted for publication (in revised form) March 12, 1992.

[†] Department of Systems Science and Mathematics, Campus Box 1040, Washington University, One Brookings Drive, Saint Louis, Missouri 63130-4899.

[‡] This author's research was partially supported by Department of Energy grant DE-FG02-90ER14140.

FIG. 1.1. Closed-loop system corresponding to (G^j, C) .

The main problem of ascertaining stability of a linear time-varying system is that the problem is not equivalent to localizing the eigenvalues of a possibly time-varying matrix to a certain region of the complex plane. In fact, a time-varying system stable in frozen time (i.e., stable for each time instance) is not necessarily a stable time-varying system (see [8] and [9]). Stability can be ascertained, however, if the parameters of the system are varying sufficiently slowly. It has been shown by Desoer [10] that there exists open regions of the parameter space with the property that if the parameter vector of the system resides in such a region for all times, then the associated time-varying system is indeed stable (see also [11]–[13]). A sufficiency criterion for stabilizing a time-varying system is, therefore, one of choosing a compensator that localizes the coefficients of the closed-loop system to within such an open region. For reasons of robustness and fault tolerance, however, one is interested in stabilizing, not just a single plant, but an entire r -tuple of plants. This problem is now described as follows.

PROBLEM 1.3 (simultaneous stabilizability problem). Given an r -tuple of linear time-varying plants G^1, G^2, \dots, G^r , find, if possible, a linear time-varying output feedback compensator C that simultaneously stabilizes each one of the closed-loop systems (G^j, C) , $j = 1, \dots, r$.

In Problem 1.3 (G^j, C) denotes the closed-loop system described in Fig. 1.1. For an introduction to the simultaneous stabilization problem for time-invariant systems, we refer to [1]–[3]. The main idea is that G^1 is a nominal plant, which, as a result of sudden component failures, may take up $r - 1$ additional different modes G^2, \dots, G^r . The design goal is to construct a compensator that stabilizes the nominal plant together with all its failed modes simultaneously. To obtain a tighter control on the response of the closed-loop system, we consider the following problem.

PROBLEM 1.4 (simultaneous coefficient assignment problem). Given an r -tuple of m input p output *autoregressive moving average* (ARMA) models of lag ℓ , denoted by $\{G^j\}$ and defined by

$$(1.1) \quad y_k + \sum_{i=1}^{\ell} D_k^j(i) y_{k-i} = \sum_{i=1}^{\ell} N_k^j(i) v_{k-i}.$$

When does there exist a compensator C of lag q , defined by

$$(1.2) \quad \begin{aligned} u_k^* + \sum_{i=1}^q \bar{D}_k(i) u_{k-i}^* &= \sum_{i=0}^q \bar{N}_k(i) y_{k-i} \\ v_k &= u_k - u_k^*, \end{aligned}$$

that will assign the coefficients $\Delta_k^j(i)$ of the equations of the closed-loop systems that (G^j, C) described as

$$(1.3) \quad \Delta_k^j(0)y_{1,k} + \sum_{i=1}^{\ell+q} \Delta_k^j(i)y_{1,k-i} = \sum_{i=0}^{\ell+q} r_k^j(i)u_{k-i}, j = 1, \dots, r.$$

The main contributions of this paper are now described. In §2 we consider a general linear input–output map and obtain a necessary and sufficient condition as to when such a map is realizable as a linear time-varying ARMA model of finite lag. We introduce left and right fraction representation of an LTV system and show via examples that, unlike the time-invariant case, the existence of one does not imply the existence of the other. However, in a suitable topology on the space of LTV ARMA systems, a generic system is shown to admit either of the two representations. In §3, we pose and analyze a simplified coefficient assignment problem. The proposed problem consists of splitting the closed-loop system into a cascade of ARMA and a moving average subsystem. The problem considered is to assign the coefficients of the autoregressive part of the ARMA subsystem. We show that the solution to the problem consists of analyzing linear equations with time-varying matrices. Under a suitable topology we consider an r -tuple of m -input p -output plants and show that a sufficient condition for the proposed coefficient assignment problem is given by $rp < m + p$. Under an additional technical condition (3.29) the above inequality is shown to be sufficient for simultaneous stabilization of the r -tuple of plants as well. In §4 we consider the closed-loop system as a single ARMA system and propose assigning the coefficients of the autoregressive part. The coefficient assignment problem is considered when $r = 1$ and m and p are arbitrary. A nonlinear time-varying iteration scheme that recursively assigns the coefficients of such time-varying systems in the closed-loop is described. Such an iterative scheme appears to be new in the literature and has no counterpart in the time-invariant system theory.

The main technique that we use in this paper is that of a time-varying version of the z -transform as part of an operational algebra for discrete-time, linear time-varying systems. Such an operational algebra has also been used previously by Kamen, Khargonekar, Poolla, and Hwang [14]–[17]. More recently a continuous-time version of the above operational algebra is being used by Tsakalis and Ioannou [18] and [19].

The main idea of this paper is to describe a time-varying version of the return difference matrix and to ensure that the coefficients of this matrix can be assigned arbitrarily under a sufficiency condition. Such a sufficiency condition in principal is a generalization of the results on “pole placement by dynamic compensation” for linear time-invariant systems (see [20]–[22]). The time-varying nature of the problem considered here imposes restrictions that did not exist in the literature concerning time-invariant systems. For example, we see in this paper that a coefficient-assigning compensator for a single-input single-output plant can be obtained by solving a nonlinear difference equation recursively. When restricted to time-invariant parameters, the difference equation reduces to the well-known linear algebraic equation of the type $Sx = b$, where S is a Sylvester matrix. Stability analysis of the proposed nonlinear difference equation has not been carried out in general and is a subject of future research.

There are other approaches [23] and [24] to stabilization, simultaneous stabilization of linear time-varying systems in the literature. For example, [23] deals with continuous-time systems wherein the input–output time-varying plant is modeled as an operator between two suitable function spaces. Among the results, it is shown that if an r -tuple of linear time-varying plants is internally stabilizable individually, then the r -tuple is simultaneously stabilizable by a stable linear time-varying compensator. The main difference between

this paper and the approach presented in [23] is now described. In this paper, we deal with linear time-varying systems in discrete time. Moreover, the time-varying system is described as a parametric variation on the space of time-invariant systems. Thus, this paper addresses problems in system design that pertain to real time adaptation of the compensator parameters as a result of real time changes in the plant parameters. The main system design problem that we consider is coefficient assignment, wherein no assumption is made about the parameters of the plants and compensators for all future times. In particular, the parameters of the plants and compensators are not assumed to be known completely. In fact, in this paper we assume that the future values of the plant parameters are unknown. Of course, to implement the coefficient assignment algorithms presented, we need to know the values of the plant parameters for an a priori fixed span of time (depending upon the lags of the systems) in the future. This adds a new twist to the problem of compensator design for a time-varying plant. Estimating the parameters of a time-varying plant in the immediate future appears to be an integral part of compensating a time-varying system, in discrete time, and to the best of our knowledge has never been considered before in the literature.

2. Representations of time-varying input–output maps. In this section, we consider a linear time-varying input–output map and study the problem of realizing the map as an impulse response of a linear time-varying *autoregressive moving average* (LTV ARMA) model of finite lag q . LTV ARMA models are of interest because the plants and compensators considered in subsequent sections of this paper are modeled as LTV ARMA systems, i.e., as ARMA systems with time-varying parameters.

Linear time-varying input–output maps are described by their impulse response sequence. We derive condition on the impulse response parameters so that it is realizable as an impulse response of an LTV ARMA system of a given lag q . To derive the realizability condition and also in later sections to describe the compensator that assigns coefficients of the closed-loop system, we find it convenient to introduce the notion of transfer function for an LTV input–output map. Such a transfer function is an obvious generalization of the z -transform methods well known in linear time-invariant discrete-time system design. To describe the transfer function, we need to introduce an operational algebra on the space of infinite power series with time-varying coefficients. Establishing connection between LTV input–output system, LTV ARMA system, and LTV transfer functions form the core of the main results described in this section.

Consider a m -input p -output LTV input–output map described by its impulse response sequence $H_j(i)$, where $H_j(i)$ is a $p \times m$ matrix defined to be the output at time j corresponding to a unit impulse at time $j - i$. To impose causality, we set $H_j(i) = 0$, the zero matrix, for all $i > j$.

Using linearity, it is clear that the impulse response sequence $H_j(i)$ completely specifies the input–output map. In fact, if u_j, u_{j-1}, \dots is a sequence of m vector inputs at time $j, j-1, \dots$, respectively, we have

$$(2.1) \quad y_j = \sum_{\ell=0}^x H_j(\ell) u_{j-\ell},$$

where y_j is the p -vector output at the time instant j . Equation (2.1) will be referred to as the LTV input–output map. The realization problem that we now consider is described as follows.

PROBLEM 2.1. Given a time-varying ARMA model of finite lag ℓ described by

$$(2.2) \quad y_k + \sum_{i=1}^{\ell} D_k(i) y_{k-i} = \sum_{i=0}^{\ell} N_k(i) u_{k-i},$$

where $D_k(i), i = 1, 2, \dots, \ell$ are $p \times p$ matrices and $N_k(i), i = 0, 1, \dots, \ell$ are $p \times m$ matrices. When is it true that the impulse response of a LTV input-output map described by (2.1) coincides with the impulse response of an ARMA model of type (2.2)?

We will see subsequently in this section that a necessary and sufficient condition for the above realization is given by a sequence of recursive conditions on the impulse response sequence H_j^i . The procedure is in principal similar to checking ranks of Hankel matrices in the theory of linear time-invariant systems.

Before we proceed to study Problem 2.1, we introduce an operational algebra and consider the notion of a transfer function for LTV input-output systems and LTV ARMA systems. This is done as follows.

Let y_k be a discrete-time vector sequence. Define a shift operator z^{-i} as follows:

$$(2.3) \quad z^{-i} y_k = y_{k-i}.$$

In the notation of (2.3), extending the operator z^{-i} linearly, we can write (2.1) as

$$(2.4) \quad y_k = \left[\sum_{\ell=0}^x H_k(\ell) z^{-\ell} \right] u_k.$$

The infinite power series

$$(2.5) \quad \mathcal{H}(z^{-1}) = \sum_{\ell=0}^x H_k(\ell) z^{-\ell}$$

is defined to be the transfer function of the LTV input-output map described by (2.1). We now define an operation of multiplication of two infinite power series of the type (2.5). Denote the multiplication operation by \circ .

Let

$$(2.6) \quad \mathcal{J}(z^{-1}) = \sum_{\ell=0}^x J_k(\ell) z^{-\ell}$$

be another infinite power series. We define

$$(2.7) \quad \mathcal{H}(z^{-1}) \circ \mathcal{J}(z^{-1}) = \sum_{\ell_1=0}^{\infty} \sum_{\ell_2=0}^x [(H_k(\ell_1) z^{-\ell_1}) \circ (J_k(\ell_2) z^{-\ell_2})],$$

where

$$(2.8) \quad (H_k(\ell_1) z^{-\ell_1}) \circ (J_k(\ell_2) z^{-\ell_2}) = H_k(\ell_1) J_{k-\ell_1}(\ell_2) z^{-(\ell_1+\ell_2)}.$$

Of course, we assume that $H_k(\ell), J_k(\ell)$ are matrices of compatible dimension so that the product $H_k(\ell_1) J_{k-\ell_1}(\ell_2)$ is defined. The following straightforward properties of the multiplication operation are now stated without proof.

PROPOSITION 2.2. *Let $\mathcal{H}(z^{-1}), \mathcal{J}(z^{-1}), \mathcal{L}(z^{-1})$ be a set of three infinite power series. Assuming that $\mathcal{H}(z^{-1}) \circ \mathcal{J}(z^{-1})$ and $\mathcal{J}(z^{-1}) \circ \mathcal{L}(z^{-1})$ are defined, we have the following:*

1. The multiplication operation \circ is associative, i.e.,

$$[\mathcal{H}(z^{-1}) \circ \mathcal{J}(z^{-1})] \circ \mathcal{L}(z^{-1}) = \mathcal{H}(z^{-1}) \circ [\mathcal{J}(z^{-1}) \circ \mathcal{L}(z^{-1})].$$

2. The multiplication operation is not commutative, i.e.,

$$\mathcal{H}(z^{-1}) \circ \mathcal{J}(z^{-1}) \neq \mathcal{J}(z^{-1}) \circ \mathcal{H}(z^{-1})$$

in general, even when the right-hand side is defined.

3. $[\mathcal{H}(z^{-1}) \circ \mathcal{J}(z^{-1})]y_k = \mathcal{H}(z^{-1})[\mathcal{J}(z^{-1})y_k]$.

We now consider the following definition.

DEFINITION 2.3 (existence of inverse). Let $\mathcal{Q}(z^{-1})$ be an infinite power series with square coefficient matrices of size $\alpha \times \alpha$. If there exists $\mathcal{W}(z^{-1})$, an infinite power series with square coefficient matrices of size $\alpha \times \alpha$ such that

$$\mathcal{Q}(z^{-1}) \circ \mathcal{W}(z^{-1}) = \mathcal{W}(z^{-1}) \circ \mathcal{Q}(z^{-1}) = I_\alpha,$$

where I_α is an identity matrix of size $\alpha \times \alpha$, then $\mathcal{W}(z^{-1})$ is called an *inverse* of $\mathcal{Q}(z^{-1})$ and we write

$$\mathcal{W}(z^{-1}) = \mathcal{Q}^{-1}(z^{-1}).$$

Not all infinite power series would have an inverse. The following proposition is important, but involves straightforward checking. Hence, the proof is omitted.

PROPOSITION 2.4. *Let*

$$\mathcal{D}(z^{-1}) = I_p + \sum_{i=1}^{\ell} D_k(i)z^{-i},$$

where $D_k(i)$ -s are $p \times p$ matrices. Then $\mathcal{D}(z^{-1})$ has an unique inverse given by

$$D^{-1}(z^{-1}) = A_k(0) + A_k(1)z^{-1} + A_k(2)z^{-2} + \cdots,$$

where

$$\begin{aligned} A_k(0) &= I_p \\ A_k(1) &= -A_k(0)D_k(1) \\ A_k(2) &= -A_k(1)D_{k-1}(1) - A_k(0)D_k(2), \text{ etc.} \end{aligned}$$

The LTV ARMA model (2.2) can be written as

$$(2.9) \quad \left[I + \sum_{i=1}^{\ell} D_k(i)z^{-i} \right] y_k = \left[\sum_{i=0}^{\ell} N_k(i)z^{-i} \right] u_k.$$

Using Proposition 2.4, we can now write

$$(2.10) \quad y_k = \left[I + \sum_{i=1}^{\ell} D_k(i)z^{-i} \right]^{-1} \left[\sum_{i=0}^{\ell} N_k(i)z^{-i} \right] u_k$$

$$(2.11) \quad = \Psi(z^{-1})u_k.$$

The power series $\Psi(z^{-1})$ is defined to be the transfer function of an LTV ARMA model of lag ℓ . We would now define the left and right representations of LTV systems as follows.

DEFINITION 2.5. The transfer function (2.5) of the input–output map (2.1) is said to have a left factorization

$$(2.12) \quad \Psi_L(z^{-1}) = \left(I + \sum_{i=1}^{\ell} D_k(i)z^{-i} \right)^{-1} \circ \left(\sum_{j=0}^{\ell} N_k(j)z^{-j} \right)$$

of lag ℓ if

$$(2.13) \quad \left(\sum_{i=0}^{\infty} H_k(i)z^{-i} \right) = \Psi_L(z^{-1})$$

and a right factorization

$$(2.14) \quad \Psi_R(z^{-1}) = \left(\sum_{j=0}^{\ell} \bar{N}_k(j)z^{-j} \right) \circ \left(I + \sum_{i=1}^{\ell} \bar{D}_k(i)z^{-i} \right)^{-1}$$

of lag ℓ , if

$$(2.15) \quad \left(\sum_{i=0}^{\infty} H_k(i)z^{-i} \right) = \Psi_R(z^{-1}).$$

It is clear from (2.9) and (2.10) that Problem 2.1 involves finding conditions under which the transfer function (2.5) admits a left factorization of lag ℓ . The following theorem completely solves the problem.

THEOREM 2.6. *The infinite power series (2.5) admits a left factorization of lag ℓ if and only if there exists $p \times p$ matrices $D_k(1), \dots, D_k(\ell)$ such that*

$$(2.16) \quad \begin{bmatrix} H_k^T(\ell+1) \\ H_k^T(\ell+2) \\ \vdots \end{bmatrix} + \begin{bmatrix} H_{k-1}^T(\ell) & H_{k-1}^T(\ell-1) & \cdots & H_{k-\ell}^T(1) \\ H_{k-1}^T(\ell+1) & H_{k-2}^T(\ell) & \cdots & H_{k-\ell}^T(2) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} D_k(1)^T \\ D_k(2)^T \\ \vdots \\ D_k(\ell)^T \end{bmatrix}.$$

The infinite power series (2.5) admits a right factorization of lag ℓ if and only if there exist $p \times p$ matrices $D_k(1), \dots, D_k(\ell)$ such that

$$(2.17) \quad [H_k^T(\ell+1)H_{k+1}^T(\ell+2)\cdots] \\ = [D_{k-\ell}^T(1)D_{k-\ell+1}^T(2)\cdots D_{k-1}^T(\ell)] \begin{bmatrix} H_k^T(\ell) & H_{k+1}^T(\ell+1) & \cdots \\ H_k^T(\ell-1) & H_{k+1}^T(\ell) & \cdots \\ \vdots & \vdots & \vdots \\ H_k^T(1) & H_{k+1}^T(2) & \cdots \end{bmatrix}.$$

Proof. We prove Theorem 2.6 for the case of left factorizations. The case of right factorizations is similar and is omitted.

The impulse response matrix (2.1) admits a left factorization of lag ℓ if and only if there exists $D_k(i), N_k(i)$ such that (2.13) is satisfied. It follows that

$$(2.18) \quad \left[\left(I + \sum_{i=1}^{\ell} D_k(i)z^{-i} \right) \circ \left(\sum_{i=0}^{\infty} H_k(i)z^{-i} \right) \right] \left(\sum_{i=0}^{\ell} N_k(i)z^{-i} \right).$$

Expanding both sides of (2.18) and equating like powers of z^{-1} yields the desired result. \square

Remark 2.7. For linear time-invariant systems, the notion of left and right factorization is well known [25]. For time-varying systems, these were introduced in [14] and [17]. The realizability condition introduced in Theorem 2.6 is new. Note that if the impulse response sequence is arranged in the form of a matrix

$$(2.19) \quad \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & H_0(0) & H_1(1) & H_2(2) & H_3(3) & H_4(4) & \cdot & \cdot \\ \cdot & \cdot & O & H_1(0) & H_2(1) & H_3(2) & H_4(3) & \cdot & \cdot \\ \cdot & \cdot & O & O & H_2(0) & H_3(1) & H_4(2) & \cdot & \cdot \\ \cdot & \cdot & O & O & O & H_3(0) & H_4(1) & \cdots & \\ \cdot & \cdot & O & O & O & O & H_4(0) & \cdots & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix},$$

the conditions for left (right) factorization can be viewed as a rank condition on the columns (rows) of the above matrix. In time-invariant system theory the above rank condition would reduce to checking ranks of a suitable Hankel matrix. The results are quite standard [26], and, therefore, the details are omitted.

Remark 2.8. It may be noted that, unlike the results in [14] and [17], wherein left and right representations for input-output maps of the form (2.1) have also been given, we do not have to construct state-space realizations of minimal order for (2.1) as an intermediary step.

The following two examples serve to illustrate the important fact that, unlike the time-invariant case, the existence of a left factorization of finite lag for a time-varying system does not imply the existence of a right factorization of finite lag and vice-versa. Both examples are for single-input single-output systems.

Example 2.9. Consider an input-output system defined by the following impulse response sequence

$$(2.20) \quad H_k(0) = 1 \quad k = 0, 1, 2, \dots$$

For $i = 1, 2, 3, \dots$

$$\begin{aligned} H_{k+i}(i) &= 1 \text{ for } k \geq 0, k \text{ even} \\ &= 0 \text{ for } k \geq 0, k \text{ odd.} \end{aligned}$$

Let $H_j(i) = 0$ for all other values of i, j . It is easy to see that (2.20) admits a left factorization of the form (2.12) of lag 1 given by $N_k(0) = 1, N_k(1) = 0, D_k(1) = -1$. On the other hand, (2.20) does not admit a right factorization of finite lag.

Example 2.10. Consider an input-output system defined by the following impulse response sequence:

$$(2.21) \quad H_k(0) = 1 \quad k = 0, 1, 2, \dots$$

For $j = 1, 2, \dots$

$$(2.22) \quad H_k(j) = 1 \quad k \geq 0, k \text{ even}$$

$$(2.23) \quad = 0 \quad k \geq 0, k \text{ odd.}$$

Let $H_j(i) = 0$ for all other values of i, j . It is straightforward to check that (2.21) admits no left factorization, but admits a right factorization of lag 1 given by

$$(1 + N_k(1)z^{-1})(1 - z^{-1})^{-1},$$

where

$$N_k(1) = -1 + H_k(1).$$

Remark 2.11. Left factorizations of finite lag always admit ARMA representations of finite lag as well as state-space realizations of finite order [14], [27].

Example 2.10 raises the concern that right factorizations, unlike left factorizations, may be difficult to implement. For this reason, we will conclude this section by showing that a right factorization

$$(2.24) \quad \mathcal{N} \circ \mathcal{D}^{-1} \equiv \left(\sum_{i=0}^{\bar{\ell}} \bar{\mathcal{N}}_k(i) z^{-i} \right) \circ \left(I + \sum_{i=1}^{\bar{\ell}} \bar{\mathcal{D}}_k(i) z^{-i} \right)^{-1}$$

of finite lag (i) generically admits a left factorization of finite lag and (ii) can always be realized in state space form.

We now consider the following topology for the space of right factorization of lag $\leq \bar{\ell}$. Note that the vector of matrices

$$(\bar{N}_k(0), \dots, \bar{N}_k(\ell), \bar{D}_k(1), \dots, \bar{D}_k(\ell)) \in \mathbb{R}^N,$$

where

$$N = \bar{\ell} p^2 + (\bar{\ell} + 1) p m$$

for each $k = 0, 1, 2, \dots$. Thus, every right factorization of the form (2.24) is a point in the product space

$$(2.25) \quad \prod_{j=0}^{\infty} \mathbb{R}^N = \mathcal{P}.$$

We now equip \mathcal{P} with the product topology (see [28]).

DEFINITION 2.12. A set \mathcal{G} of right factorizations is said to be generic if \mathcal{G} can be written as an intersection of a countable number of open and dense sets in \mathcal{P} .

The ARMA realization of a system of the form (2.24) is given by the following theorem.

THEOREM 2.13. *Consider a generic element in the space of right factorizations of lag $\leq \bar{\ell}$ with m inputs and p outputs. There always exists a left factorization of lag ℓ , where ℓ is the smallest integer satisfying $\ell p \geq \bar{\ell} m$ such that the two factorizations correspond to the same infinite power series for all $k \geq \ell$.*

Proof. See Appendix I.

Note 2.14. The basic interpretation of Theorem 2.13 is that almost all LTV transfer functions with a right factorization also have a left factorization and, therefore, can be realized as an LTV ARMA system. Of course, we could also define a generic set of left factorizations to show that generically a transfer function has left factorization (right factorization) if it has a right factorization (respectively, left factorization).

We will now state that right factorizations of finite lag (2.24) can always be realized in state-space form. This realization has not been used subsequently in this paper. It is stated only to satisfy our curiosity that even though a right factorization may not have a left factorization, it can still be realized as a state-space system, but possibly not as an ARMA system. The result follows with modifications from [14] and [16].

THEOREM 2.15 (state-space realization). *The right factorization given by (2.24) can always be realized as an ℓ th-order state-space system*

$$\begin{aligned}x(k+1) &= F(k)x(k) + G(k)u(k); x(0) = 0 \\ y(k) &= H(k)x(k) + J(k)u(k),\end{aligned}$$

where

$$(2.26) \quad F(k) \equiv \begin{bmatrix} O & O & O & O & -\bar{D}_k(\ell) \\ I & O & O & O & -\bar{D}_{k-1}(\ell-1) \\ O & I & O & O & -\bar{D}_{k-2}(\ell-2) \\ & & \dots & \vdots & \vdots \\ O & O & I & O & -\bar{D}_{k-\ell+2}(2) \\ O & O & O & I & -\bar{D}_{k-\ell+1}(1) \end{bmatrix},$$

$$(2.27) \quad G \equiv \text{col} [I, O, O, \dots, O],$$

$$(2.28) \quad H(k) \equiv [W_k(1), W_k(2), \dots, W_k(\ell-1), W_k(\ell)],$$

and

$$(2.29) \quad J(k) \equiv W_k(0),$$

where $W_k(i)$ is the coefficient of z^{-i} in the formal power series expansion of (2.24).

Proof. We refer the interested reader to [14] and [16].

The main contributions of this section are now summarized. Starting from the impulse response sequence of an input–output map, we introduce a formal infinite power series. We then completely answer the question as to when such a power series can be represented as a left (right) factorization. Existence of a left factorization enables us to construct an LTV ARMA system that realizes the impulse response sequence. Right factorization, on the other hand, can be realized as a state-space system. This fact is of independent interest, but is not used subsequently in this paper. Finally, we show that the existence of left (right) factorization in general does not imply the existence of respectively right (left) factorization, although, for a generic transfer function, that is indeed the case. This fact should be contrasted with linear time-invariant transfer functions, wherein existence of one implies the existence of the other.

3. A nonrecursive compensator design technique for simultaneous coefficient assignment. In this section, we shall consider Problem 1.4 regarding simultaneous coefficient assignment of an r -tuple of m input p output LTV ARMA systems by a single LTV ARMA compensator. We also show that the coefficient assignment problem can be used to analyze Problem 1.3 as well.

To introduce the problem, let us consider the r -tuple of plants defined in (1.1). Assume that the transfer function of the j th plant is given by

$$(3.1) \quad G^j(z^{-1}) = \mathcal{D}^{j-1}(z^{-1})\mathcal{N}^j(z^{-1}),$$

where

$$(3.2) \quad \begin{aligned} \mathcal{D}^j(z^{-1}) &= I + \sum_{i=1}^{\ell} D_k^j(i)z^{-i}, \\ \mathcal{N}^j(z^{-1}) &= \sum_{i=1}^{\ell} N_k^j(i)z^{-i}, \end{aligned}$$

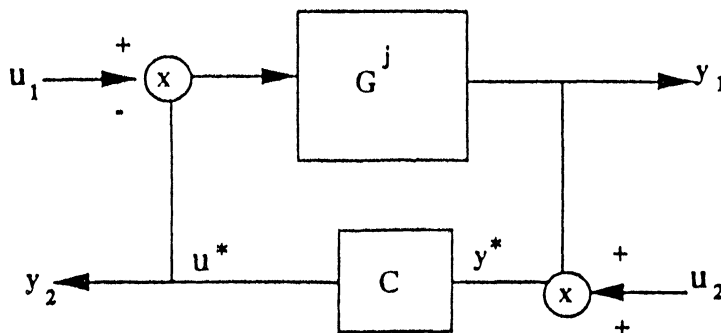


FIG. 3.1. A two-input two-output configuration of the closed-loop system.

$j = 1, 2, \dots, r$. Furthermore, consider a compensator with transfer function

$$(3.3) \quad \mathcal{C}(z^{-1}) = \bar{\mathcal{N}}(z^{-1})\bar{\mathcal{D}}^{-1}(z^{-1}),$$

where

$$(3.4) \quad \begin{aligned} \bar{\mathcal{D}}(z^{-1}) &= I + \sum_{i=1}^q \bar{D}_k(i)z^{-i}, \\ \bar{\mathcal{N}}(z^{-1}) &= \sum_{i=0}^q \bar{N}_k(i)z^{-i}. \end{aligned}$$

It may be noted that the compensator (3.3) may not have a left representation, and therefore, may not have an LTV ARMA realization unless it satisfies the generic conditions of Theorem 2.13. Assume that the plants and the compensator are put in the configuration given by Fig. 3.1.

The following is the transfer function of the closed-loop system with respect to the j th plant.

$$(3.5) \quad \begin{bmatrix} y_{1,k} \\ y_{2,k} \end{bmatrix} = \begin{bmatrix} (\bar{\mathcal{D}} \circ (\mathcal{D}^j \circ \bar{\mathcal{D}} + \mathcal{N}^j \circ \bar{\mathcal{N}})^{-1} \circ \mathcal{N}^j) & (-I + \bar{\mathcal{D}} \circ (\mathcal{D}^j \circ \bar{\mathcal{D}} + \mathcal{N}^j \circ \bar{\mathcal{N}})^{-1} \circ \mathcal{D}^j) \\ (\bar{\mathcal{N}} \circ (\mathcal{D}^j \circ \bar{\mathcal{D}} + \mathcal{N}^j \circ \bar{\mathcal{N}})^{-1} \circ \mathcal{N}^j) & (\bar{\mathcal{N}} \circ (\mathcal{D}^j \circ \bar{\mathcal{D}} + \mathcal{N}^j \circ \bar{\mathcal{N}})^{-1} \circ \mathcal{D}^j) \end{bmatrix} \cdot \begin{bmatrix} u_{1,k} \\ u_{2,k} \end{bmatrix} \quad j = 1, 2, \dots, r.$$

Note that in (3.5), $\bar{\mathcal{D}}, \bar{\mathcal{N}}, \mathcal{D}^j, \mathcal{N}^j$ stands for $\bar{\mathcal{D}}(z^{-1}), \bar{\mathcal{N}}(z^{-1}), \mathcal{D}^j(z^{-1}), \mathcal{N}^j(z^{-1})$ defined in (3.2) and (3.4). Whenever convenient, in the future we will suppress z^{-1} . We now consider the following simultaneous coefficient assignment problem.

PROBLEM 3.1 (coefficient assignment problem). Given an r -tuple of plants described by (3.1), find a compensator of the type (3.3) such that

$$(3.6) \quad \mathcal{D}^j \circ \bar{\mathcal{D}} + \mathcal{N}^j \circ \bar{\mathcal{N}} = I + \sum_{i=1}^{\ell+q} \Delta_k^j(i)z^{-i}$$

for a prespecified set of coefficients $\Delta_k^j(i), j = 1, \dots, r; i = 1, \dots, \ell + q$.

For time-invariant systems, the quantity in the left-hand side of (3.6) is the *return difference*, determinant of which is the closed-loop characteristic polynomial. Thus, Problem 3.1 is the analogue of the pole placement problem for time-invariant systems.

It is not a priori clear how the coefficients $\Delta_k^j(i)$ are related to the input–output response of the closed-loop system (3.5). In particular, we might ask the following question: *If a plant is coefficient assignable by some choice of a compensator, is the closed loop-system stable in the bounded input bounded output sense?* For a time-invariant system, the answer to this question is always affirmative. For a time-varying system, the BIBO stability is not necessarily guaranteed. To ascertain the BIBO stability of the closed-loop system, the compensator coefficients have to be uniformly bounded. To examine this question, let us consider the transfer function between $y_{1,k}$ and $u_{1,k}$ for the j th plant given by

$$(3.7) \quad y_{1,k} = [\bar{\mathcal{D}} \circ (\mathcal{D}^j \circ \bar{\mathcal{D}} + \mathcal{N}^j \circ \bar{\mathcal{N}})^{-1} \circ \mathcal{N}^j] u_{1,k},$$

which may be written as a cascade of two interconnected subsystems given by

$$(3.8) \quad y_{1,k} = \bar{\mathcal{D}} u_{1,k}^*,$$

$$(3.9) \quad [\mathcal{D}^j \circ \bar{\mathcal{D}} + \mathcal{N}^j \circ \bar{\mathcal{N}}] u_{1,k}^* = \mathcal{N}^j u_{1,k}.$$

Clearly, (3.7) is BIBO stable if each of the two subsystems (3.8) and (3.9) are BIBO stable. If we assume that the coefficients of the plants $D_k^j(i)$ and $N_k^j(i)$ are bounded uniformly in k , then an important question to ask is whether or not Problem 3.1 can be solved by a compensator with coefficients $\bar{D}_k(i)$, $\bar{N}_k(i)$, bounded uniformly in k . We therefore consider the following problem.

PROBLEM 3.2 (bounded coefficient assignment problem). Given an r -tuple of plants described by (3.1), find a compensator with coefficients $\bar{D}_k(i)$, $\bar{N}_k(i)$ uniformly bounded in time k such that (3.6) is satisfied for a prespecified set of coefficients $\Delta_k^j(i)$.

Note that if N^j is uniformly bounded, then for an appropriate choice of $\Delta_k^j(i)$, the input–output system (3.9) can be made BIBO stable provided that the coefficients $\Delta_k^j(i)$ are assignable. This fact follows easily from Desoer [10] and has been subsequently studied in detail by Bouthellier [27]. The basic idea is to choose $\Delta_k^j(i)$ such that they are slowly varying in between any two consecutive times. Following [10] and [27], we could construct a chain of open neighborhood Ω_k in the space of coefficients such that for all k , we have

$$(\Delta_k^j(1), \Delta_k^j(2), \dots, \Delta_k^j(\ell + q)) \in \Omega_k.$$

For the above choice of coefficients, Problem 3.2 would guarantee simultaneous BIBO stabilizability of the r -tuple of plants. The main result of this section is described below.

THEOREM 3.3. *A generic r -tuple of $p \times m$ plant is coefficient assignable if and only if*

$$(3.10) \quad p + m > rp.$$

Furthermore, if (3.10) is satisfied, then the r -tuple is coefficient assignable by a compensator of lag q where q is the smallest integer satisfying

$$(3.11) \quad q[m + p - rp] \geq rp\ell - m,$$

where ℓ is the lag of each of the r plants.

To get an idea as to how tight the bounds (3.10) and (3.11) are, we consider the following theorem.

THEOREM 3.4. *A generic r -tuple of $p \times m$ plants can be assigned with bounded coefficients uniformly for all k and for all plants in the generic set by some choice of feedback*

compensator (where the compensator can depend on the choice of the r -tuple of plants) if and only if (3.10) is satisfied.

Thus, for a generic set of r -tuple of plants, if (3.10) is not satisfied, then not only is it not possible to assign coefficients simultaneously, but it is also not possible to restrict the coefficients to a bounded set uniformly in k for all plants in a generic set. We now consider the proofs of Theorems 3.3 and 3.4.

Proof of Theorem 3.3. Consider the r -tuple of plants (3.1) together with the compensator (3.3). In the notation described by (3.2), (3.4) we can equate the like powers of z^{-1} in (3.6) to obtain the following linear equations:

$$(3.12) \quad M_k^j \nu_k = \Delta_k^j,$$

where

$$(3.13) \quad \nu_k = \text{col} [I, \bar{D}_{k+1}(1), \bar{D}_{k+2}(2), \dots, \bar{D}_{k+q}(q), \bar{N}_k(0), \bar{N}_{k+1}(1), \dots, \bar{N}_{k+q}(q)],$$

$$(3.14) \quad \Delta_k^j = \text{col} [\Delta_{k+1}^j(1), \Delta_{k+2}^j(2), \dots, \Delta_{k+\ell+q}^j(\ell+q)],$$

$$M_k^j =$$

$$(3.15) \quad \begin{bmatrix} D_{k+1}^j(1) & I & O & N_{k+1}^j(1) & O & O \\ D_{k+2}^j(2) & D_{k+2}^j(1) & \vdots & N_{k+2}^j(2) & N_{k+2}^j(1) & \vdots \\ \vdots & D_{k+3}^j(2) & O & \vdots & N_{k+3}^j(2) & \vdots \\ \cdot & \cdot & \dots & I & \cdot & O \\ D_{k+\ell}^j(\ell) & \vdots & D_{k+q+1}^j(1) & N_{k+\ell}^j(\ell) & \vdots & N_{k+q+1}^j(1) \\ O & D_{k+\ell+1}^j(\ell) & D_{k+q+2}^j(2) & O & N_{k+\ell+1}^j(\ell) & N_{k+q+2}^j(2) \\ O & O & \vdots & \vdots & O & \cdot \\ O & \vdots & \cdot & \cdot & \vdots & \vdots \\ O & O & D_{k+q+\ell}^j(\ell) & O & O & N_{k+q+\ell}^j(\ell) \end{bmatrix}$$

for $j = 1, 2, \dots, r$. If we now define the matrix

$$(3.16) \quad M_k = \text{col} (M_k^1, M_k^2, \dots, M_k^r)$$

and the matrix

$$(3.17) \quad \Delta_k = \text{col} (\Delta_k^1, \Delta_k^2, \dots, \Delta_k^r)$$

we can combine the r linear equations (3.12) as

$$(3.18) \quad M_k \nu_k = \Delta_k.$$

It is easy to check that M_k is a $rp(\ell+q) \times (q+1)(m+p)$ matrix, ν_k is a $(q+1) \times (m+p) \times p$ matrix, and Δ_k is a $rp(\ell+q) \times p$ matrix. It follows that given M_k and Δ_k we can solve (3.18) for a suitable ν_k if and only if

$$(3.19) \quad rp(\ell+q) \leq (q+1)(m+p) - p,$$

and M_k is of full column rank for each k . The inequality (3.19) follows from the requirement that (3.18) is solvable if and only if the matrix M_k after deleting the first p columns has more columns than rows. Note that the inequality (3.19) is the same as the inequality (3.11). The proof of this theorem is now complete by noting that generically the rows of M_k^* obtained from M_k by deleting the first p columns are all independent. The proof of this last statement is technical, and we refer to [27] for details. The basic idea is to choose a minor for M_k with nonidentically vanishing determinant. \square

Proof of Theorem 3.4. The sufficiency part of Theorem 3.2 follows from Theorem 3.1. We now have to prove the necessity part.

Let λ be a variable that takes on values $1, 2, 3, \dots$. Assume that the r -tuple of plants (3.1) is generically coefficient assignable by the compensator (3.3) and assume that the coefficients $\Delta_k^j(i)$ are all bounded with respect to some matrix norm; i.e., there exists $M > 0$ such that

$$(3.20) \quad \|\Delta_k^j(i)\| \leq M$$

for all $j = 1, \dots, r; i = 1, 2, \dots, l + q$ and $k = 0, 1, \dots$. Let \mathcal{P}^r be the space of r -tuples of plants equipped with the product topology similar to that described in (2.25). We now describe a map Ψ_λ for each λ given by

$$(3.21) \quad \Psi_\lambda : \mathcal{P}^r \rightarrow \mathcal{P}^r$$

described as

$$\begin{aligned} D_k^j(i) &\mapsto \lambda^{-i} D_k^j(i), \\ N_k^j(i) &\mapsto \lambda^{-i} N_k^j(i). \end{aligned}$$

It follows that Ψ_λ maps a generic set of r -tuples of plants to a generic set of r -tuples of plants. For each λ we now define the compensator

$$C_\lambda(z^{-1}) = \bar{N}_\lambda(z^{-1})\bar{D}_\lambda^{-1}(z^{-1}),$$

where

$$\begin{aligned} (3.22) \quad \bar{D}_\lambda(z^{-1}) &= I + \sum_{i=1}^q \lambda^{-i} \bar{D}_k(i) z^{-i}, \\ \bar{N}_\lambda(z^{-1}) &= \sum_{i=0}^q \lambda^{-i} \bar{N}_k(i) z^{-i}. \end{aligned}$$

Thus, we conclude that for each λ , there exists an open and dense set S_λ such that every r -tuple of plants in S_λ can be assigned with coefficients $\Delta_k^j(i)$, by a compensator of type (3.22), such that

$$(3.23) \quad \|\lambda^i \Delta_k^j(i)\| \leq M$$

for $i = 1, 2, \dots, \ell + q$. Define

$$(3.24) \quad U = \cap_{\lambda=1}^x S_\lambda.$$

Thus, for every r -tuple of plants in U , there is a sequence of compensators such that the corresponding closed-loop system has coefficients in an arbitrary small neighborhood of 0. However, the map from the space of compensators to the space of coefficients is a linear

map described by (3.18). It follows that the image of this linear map is closed. Hence, for every r -tuple of plants in U , there exists a compensator that places the coefficients $\Delta_k^j(i)$ at 0. In other words, we can solve the system of equation

$$(3.25) \quad M_k \nu_k = 0.$$

The proof of Theorem 3.4 is now completed by showing that there exist open sets of r -tuples of plants for which (3.25) is not satisfied if $rp \geq p + m$. This is done as follows.

Define

$$(3.26) \quad \begin{aligned} D_j^*(i) &= \text{col} [D_j^1(i), D_j^2(i), \dots, D_j^r(i)], \\ N_j^*(i) &= \text{col} [N_j^1(i), N_j^2(i), \dots, N_j^r(i)]. \end{aligned}$$

We now make specific choices of $D_j^*(i), N_j^*(i)$ as follows. As $m + p \leq rp$, which implies $rp\ell > m$, we set

(i) the $rp\ell \times m$ matrix

$$(3.27) \quad \text{col} [N_{k+1}^*(1), N_{k+2}^*(2), \dots, N_{k+\ell}^*(\ell)] \equiv [e_1, e_2, \dots, e_m],$$

where e_i is the i th standard basis vector in $\mathbb{R}^{rp\ell}$, $i = 1, \dots, m$;

(ii) the first column of the matrix

$$\text{col} [-D_{k+1}^*(1), -D_{k+2}^*(2), \dots, -D_{k+\ell}^*(\ell)]$$

to be e_{m+1} , where e_{m+1} is the $m+1$ st standard basis vector in $\mathbb{R}^{rp\ell}$; and

(iii) the $rp \times (p + m)$

$$(3.28) \quad [D_{k+\ell+j}^*(\ell), N_{k+\ell+j}^*(\ell)] \equiv [e_1, e_2, \dots, e_{m+p}],$$

where e_i is the i th standard basis vector in \mathbb{R}^{rp} , $i = 1, \dots, m + p$ for $j = 1, \dots, q$.

For the above choice of r -tuple of plants, it can be shown that (3.25) cannot be solved (see [27] for details). Furthermore, in any neighborhood of the coefficient space of the above r -tuple of plants, (3.25) has no solution. This concludes the proof. \square

Remark. The proof of Theorem 3.4 is an adaptation of a technique due originally to Anderson and Byrnes [29].

Remark 3.5. We now state and prove a result that addresses Problem 3.2.

THEOREM 3.6. *A bounded set S of r -tuple of $p \times m$ plants is coefficient assignable simultaneously by a compensator with bounded coefficients if (3.10) is satisfied and if*

$$(3.29) \quad \det [M_k M_k^T] > \epsilon$$

for some $\epsilon > 0$, which is independent of k , and for all r -tuples of plants in S .

Note that as a result of condition (3.29), S fails to be a generic set. The proof of Theorem 3.6 follows from the following simple and well-known proposition.

PROPOSITION 3.7. *Let z be an m vector and A be an m by n matrix of rank m . The n vector x such that $Ax = z$ and $x^T x$ is minimum is given by*

$$(3.30) \quad x = A^T (AA^T)^{-1} z$$

For a proof of the above proposition see Brockett [30, p. 127].

Proof of Theorem 3.6. Our basic problem is to solve (3.18) for a uniformly bounded ν_k . Clearly, in view of Proposition 3.7, the solution

$$(3.31) \quad \nu_k^* = M_k^T (M_k M_k^T)^{-1} \Delta_k$$

has the property that the columns of ν_k^* have minimum norm. Since the r -tuple of plants and the coefficients Δ_k are all bounded uniformly in k , it follows that under the assumption (3.29), ν_k^* is uniformly bounded as well. \square

To conclude this section, we reiterate the three important questions that we address in this section.

(a) For a generic set of r -tuple of $p \times m$ plants, when is it possible to coefficient assign simultaneously?

(b) For a generic set of r -tuple of $p \times m$ plants, when is it possible to assign a bounded set of coefficients simultaneously?

(c) For a set of r -tuple of $p \times m$ plants, when is it possible to assign coefficients simultaneously by a compensator with coefficients bounded uniformly in k ?

4. A recursive formulation of the coefficient assignment problem. We begin this section with the remark that, so far in this paper, Problem 1.4 has not been considered. Instead, in § 3, the closed-loop system was decomposed into ARMA and moving average subsystems. The design problem considered has been to assign the coefficients of the ARMA subsystem while maintaining an uniform bound on the coefficients of the compensators. Such a design problem leads to a simplified algorithm. To implement the algorithm, we need to solve linear equations.

We now consider Problem 1.4 for a single plant. The case for an r -tuple of plants is analogous and is not described in detail. Assume that the plants and the compensator are put in the configuration given by Fig. 3.1. For simplicity, we only consider the transfer function between y_1 and u_1 . However, unlike that in § 3, we will not decompose the transfer function (3.7) into a cascade of two transfer functions (3.8), (3.9). We will see shortly in this section that this introduces new problems, namely, the compensator parameters are not obtained by solving static linear equations one for each time. In general, Problem 1.4 reduces to a nonlinear discrete iteration on the parameter space of compensators. The algorithm, although more complicated, iteratively solves this coefficient assignment problem.

Consider the transfer function (3.7) for a single plant (i.e., assume $j = 1$). Define

$$(4.1) \quad \mathcal{X}(z^{-1}) = \sum_{i=0}^q X_k(i) z^{-i},$$

$$(4.2) \quad \Delta(z^{-1}) = \sum_{i=0}^{\ell+q} \Delta_k(i) z^{-i}$$

such that

$$(4.3) \quad \bar{\mathcal{D}} \circ (\mathcal{D} \circ \bar{\mathcal{D}} + \mathcal{N} \circ \bar{\mathcal{N}})^{-1} = \Delta^{-1} \circ \mathcal{X}.$$

The transfer function (3.7) can be written as

$$(4.4) \quad \Delta(z^{-1}) y_k = \mathcal{X}(z^{-1}) \circ \mathcal{N}(z^{-1}) u_k$$

PROBLEM 4.1 (the coefficient assignment problem). Given $\mathcal{N}(z^{-1})$ and $\mathcal{D}(z^{-1})$, find, if possible, an $\bar{\mathcal{N}}(z^{-1})$, $\bar{\mathcal{D}}(z^{-1})$ such that $\Delta_k(i)$, $i = 1, \dots, \ell+q$ can be assigned a prespecified set of coefficients.

Problem 4.1 can be stated equivalently by rewriting (4.3) as

$$(4.5) \quad \mathcal{X} \circ (\mathcal{D} \circ \bar{\mathcal{D}} + \mathcal{N} \circ \bar{\mathcal{N}}) = \Delta \circ \bar{\mathcal{D}}$$

To solve (4.5) we equate like powers of z^{-1} , $i = 0, 1, \dots, \ell + 2q$ for all $k \geq i$ and solve for $\bar{\mathcal{D}}$ and $\bar{\mathcal{N}}$. This will be accomplished by considering two sets of equations.

(A) The set of equations derived by equating like powers of z^{-1} , $i = \ell + q, \dots, \ell + 2q$ in (4.5), and

(B) The set of equations derived by equating like powers of z^{-1} , $i = 0, \dots, \ell + q - 1$ in (4.5).

Using the above two sets of equations, we derive an iterative scheme that will allow us to solve (4.5). From the set of equations (A) as described above, we get the matrix equation

$$(4.6) \quad S_k \phi_k = \psi_k,$$

where S_k is a $(q+1)p \times (q+1)p$ matrix defined by

$$(4.7) \quad \begin{bmatrix} \zeta_{k+\ell+q-1}^T(\ell+q) & \zeta_{k+\ell+q-2}^T(\ell+q-1) & \zeta_{k+\ell+q-3}^T(\ell+q-2) & \cdots & \zeta_{k+\ell}^T(\ell+1) & \zeta_{k+\ell-1}^T(\ell) \\ 0 & \zeta_{k+\ell+q-2}^T(\ell+q) & \zeta_{k+\ell+q-3}^T(\ell+q-1) & \cdots & \zeta_{k+\ell}^T(\ell+2) & \zeta_{k+\ell-1}^T(\ell+1) \\ 0 & 0 & \zeta_{k+\ell+q-3}^T(\ell+q) & \cdots & \zeta_{k+\ell}^T(\ell+3) & \zeta_{k+\ell-1}^T(\ell+2) \\ & & \vdots & & & \\ 0 & 0 & 0 & 0 & \zeta_{k+\ell}^T(\ell+q) & \zeta_{k+\ell-1}^T(\ell+q-1) \\ 0 & 0 & 0 & 0 & 0 & \zeta_{k+\ell-1}^T(\ell+q) \end{bmatrix}$$

and where we define

$$(4.8) \quad \mathcal{D}_k \circ \bar{\mathcal{D}}_k + \mathcal{N}_k \circ \bar{\mathcal{N}}_k = \sum_{i=0}^{\ell+q} \zeta_k(i) z^{-i}.$$

Moreover, ϕ_k and ψ_k are defined as follows. Note that ϕ_k is a $(q+1)p \times p$ matrix given by

$$(4.9) \quad \phi_k = \text{col} [X_{k+\ell+q-1}^T(0), X_{k+\ell+q-1}^T(1), \dots, X_{k+\ell+q-1}^T(q-1), X_{k+\ell+q-1}^T(q)]$$

and ψ_k is a $(q+1)p \times p$ matrix given by

$$(4.10) \quad \psi_k = \text{col} \left[\sum_{j=0}^q \bar{D}_{k+q-j-1}^T(q-j) \Delta_{k+\ell+q-1}^T(\ell+j), \right. \\ \cdot \sum_{j=0}^{q-1} \bar{D}_{k+q-j-2}^T(q-j) \Delta_{k+\ell+q-1}^T(\ell+j+1), \dots, \\ \cdot \sum_{j=0}^{q-i} \bar{D}_{k+q-j-i-1}^T(q-j) \Delta_{k+\ell+q-1}^T(\ell+j+i), \dots, \bar{D}_{k-1}^T(q) \Delta_{k+\ell+q-1}^T(\ell+q) \left. \right].$$

Similarly, from the set of equations (B) we obtain

$$(4.11) \quad M_k \nu_k = O,$$

where ν_k is defined as follows:

$$\nu_k = \text{col} [\bar{D}_k(0), \bar{D}_{k+1}(1), \dots, \bar{D}_{k+q}(q); \bar{N}_k(0), \bar{N}_{k+1}(1), \dots, \bar{N}_{k+q}(q)].$$

The matrix M_k is a $(\ell + q)p \times (q + 1)(p + m)$ matrix that can be shown to be a function of x_k, Δ_k and the plant parameters. If we assume that

$$(q + 1)m > p(\ell - 1),$$

it follows that (4.11) can be solved for a nonzero solution.

Using (4.6) and (4.11), we are now in a position to solve (4.5), and hence, the coefficient assignment problem, in an iterative fashion. However, before we consider the following coefficient assignment algorithm, for the sake of clarity, we will provide an overview of the basic idea. First, we will initialize the algorithm by choosing suitable (timewise) values of the plant, compensator, and parameters to be assigned (i.e., the $\Delta_j(i)$). Having obtained these values, we are then able to solve for the next set (timewise) of compensator parameters ν_k via (4.11). These values of ν_k are then used in (4.6) to solve for the next set of $X_k(i) i = 0, \dots, q$. These values of X_k are then substituted into (4.11) to solve for ν_{k+1} , which is used in turn to solve for X_{k+1} etc. ... It will be assumed that in the following coefficient assignment algorithm (4.6) and (4.11) admit solutions for all times k and that $\det \bar{D}_k(0) \neq 0$ for all k so that \bar{D}^{-1} always exists.

The coefficient assignment algorithm

Step I (initialize the algorithm). Choose values for

- (i) $\bar{D}_j(i), \bar{N}_j(i) i = 0, \dots, q, j \leq i - 1$
- (ii) $D_j(i), N_j(i) i = 0, \dots, \ell - 1, j \leq q + i - 1; i = \ell, j \leq \ell + q - 1$
- (iii) $\Delta_j(i) i = 0, 1, \dots, \ell - 1, j \leq q + i; i = \ell, \dots, \ell + q, j \leq \ell + q - 1$

Step II. Solve (4.6) for $X_j(i) i = 0, \dots, q, j = 0, 1, \dots, \ell + q - 1$ and

Step III. Using the values of $X_j(i)$ computed in Step II compute ν_0 using (4.11).

Step IV. Set $k = 0$

Step V. Obtain an estimate of the future values of the plant parameters

(1) $D_{k+q+i+1}(i), N_{k+q+i+1}(i), i = 0, 1, \dots, \ell - 1$ and $D_{k+q+\ell}(\ell), N_{k+q+\ell}(\ell)$ and choose values for

(2) $\Delta_{k+q+i+1}(i) i = 0, 1, \dots, \ell - 1$ and $\Delta_{k+q+\ell}(i) i = \ell, \dots, \ell + q$

Step VI. Solve (4.6) for $X_{k+\ell+q}(i) i = 0, \dots, q$

Step VII. Solve (4.11) for ν_{k+1}

Step VIII. Set $k = k + 1$ and return to Step V.

Remark 4.2. The values required in Step I could be based on the available knowledge of the plant parameters $D_k(i), N_k(i) i = 0, \dots, \ell$ at time $k = 0$ and the values of $\Delta_k(i) i = 0, 1, \dots, \ell + q$, which have been specified. It should also be noted that the lag q of the compensator computed via the above algorithm is the smallest nonnegative integer, which satisfies $q > [p(\ell - 1) - m]/m$.

Remark 4.3. It can be shown that, using techniques similar to those of §3, that the above algorithm can be extended to simultaneously coefficient assign an r -tuple of $p \times m$ systems, where $r < m/p + 1$. We will not elaborate on this further and refer the interested reader to [31].

We now consider two illustrative examples of the above coefficient assignment algorithm.

Example 4.4. Consider the closed-loop system given by Fig. 1.1 (assume $j = 1$) where the plant G is given by

$$y_k + a_k y_{k-1} = b_k u_k + c_k u_{k-1}$$

and the compensator C is given by the gain feedback

$$u_k^* = \bar{d}_k^{-1} y_k.$$

Writing the plant as

$$y_k = (d_k^{-1} \circ n_k) u_k,$$

where

$$d_k(z^{-1}) = 1 + a_k z^{-1}; n_k(z^{-1}) = b_k + c_k z^{-1},$$

we obtain by Theorem 3.3 the equation of the closed-loop system as follows:

$$(4.12) \quad ([d_k \circ \bar{d}_k + n_k] \circ \bar{d}_k^{-1}) y_k = n_k u_k.$$

Writing

$$(4.13) \quad (d_k \circ \bar{d}_k + n_k) \circ \bar{d}_k^{-1} = f_k^{-1} \circ (1 + e_k z^{-1}),$$

(4.12) reduces to

$$y_k + e_k y_{k-1} = f_k b_k u_k + f_k c_k u_{k-1}.$$

From (4.13) it follows that

$$f_k \circ (d_k \circ \bar{d}_k + n_k) = (1 + e_k z^{-1}) \circ \bar{d}_k,$$

i.e., equating like powers of z^{-1}

$$(4.14) \quad f_k(\bar{d}_k + b_k) = \bar{d}_k \quad \forall k \geq 0$$

and

$$(4.15) \quad f_k(a_k \bar{d}_{k-1} + c_k) = e_k \bar{d}_{k-1} \quad \forall k \geq 1.$$

Eliminating \bar{d}_k in (4.14) and (4.15), we obtain

$$(4.16) \quad f_{k+1} = \frac{f_k b_k e_{k+1}}{c_{k+1} - f_k [c_{k+1} - a_{k+1} b_k]} \quad \forall k \geq 0$$

and

$$(4.17) \quad \bar{d}_k = \frac{f_k b_k}{1 - f_k} \quad \forall k \geq 0.$$

Given the plants parameters a_k, b_k , and c_k , and given e_k , the coefficient of the closed-loop system to be assigned, (4.16) describes a nonlinear recursion in f_k . Equation (4.17), on the other hand, is a nonlinear function that computes the feedback gain in real time.

Among several questions that we might ask about (4.16) and (4.17), an important one in terms of understanding the properties of the coefficient assignment algorithm is the following.

Question 4.5. If a_k, b_k, c_k , and e_k are time invariant and given, respectively, by a, b, c , and e , what is the asymptotic behavior of (4.16) and (4.17)?

Defining $\alpha = be$, $\beta = c$, and $\gamma = -[c - ab]$, we obtain the following recursion from (4.16):

$$(4.18) \quad f_{k+1} = \frac{\alpha f_k}{\beta + f_k \gamma}.$$

Note that (4.18) has two stationary points 0 and $(\alpha - \beta)/\gamma$. It is easy to show that the second stationary point corresponds to the time-invariant solution; i.e., the corresponding value of the gain \bar{d} equals the value of the gain, which assigns the parameter e in the closed-loop system if a, b, c , and e were time invariant and known. The stationary point 0, on the other hand, corresponds to an infinite gain.

To examine the trajectory of f_{k+1} as defined by (4.18), we write

$$(4.19) \quad f_k = \frac{g_k}{h_k}$$

Substituting (4.19) into (4.18) yields

$$\frac{g_{k+1}}{h_{k+1}} = \frac{\alpha g_k}{\beta h_k + \gamma g_k},$$

which may be rewritten as

$$\begin{bmatrix} g_{k+1} \\ h_{k+1} \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ \gamma & \beta \end{bmatrix} \begin{bmatrix} g_k \\ h_k \end{bmatrix}.$$

With an initial (nonzero) estimate f_0 of the true value of $f = (\alpha - \beta)/\gamma$ we may write

$$\begin{bmatrix} g_k \\ h_k \end{bmatrix} = -f_0 \alpha^k \begin{bmatrix} \alpha - \beta \\ \gamma \end{bmatrix} + [(\beta - \alpha) + \gamma f_0] \beta^k \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Therefore, if $|\alpha| > |\beta|$ (i.e., $|be| > |c|$)

$$\lim_{k \rightarrow \infty} f_k = \lim_{k \rightarrow \infty} \frac{g_k}{h_k} = \frac{\alpha - \beta}{\gamma}.$$

On the other hand, if $|\alpha| < |\beta|$,

$$\lim_{k \rightarrow \infty} f_k = 0.$$

We may summarize the above results as follows: for choices of a, b, c , and e for which $|be| < |c|$, the adaptive gain \bar{d} tends toward 0, the infinite gain. For choices of a, b, c , and e for which $|be| > |c|$, the adaptive gain converges to the unique time-invariant solution.

Remark 4.6. Example 4.4 illustrates that under suitable conditions on the coefficients of a time-invariant plant, the coefficient assignment algorithm may be viewed as a globally convergent adaptive controller. It may also be noted that when $\alpha = \beta = 1$ or when $\alpha = \beta = -1$, $|f_k|$ converges to 0 as k tends to infinity. On the other hand, when $\alpha = -\beta = 1$, f_k is periodic of period 2.

Example 4.7. Consider the closed-loop system defined by Fig. 1.1 (assume $j = 1$), where the plant G is given by

$$(4.20) \quad y_k + d_k(1)y_{k-1} + d_k(2)y_{k-2} = n_k(1)u_{k-1} + n_k(2)u_{k-2}$$

and the compensator C is given by

$$(4.21) \quad u_k^* + \bar{d}_k(1)u_{k-1}^* = \bar{n}_k(0)y_k + \bar{n}_k(1)y_{k-1}.$$

Define $\mathcal{X}_k(z^{-1})$ given by (4.1)–(4.3) as follows:

$$\mathcal{X}_k(z^{-1}) = 1 + X_k z^{-1}.$$

It can be shown that by writing the parameters of the compensator (4.21) in terms of the plant parameters and the $\Delta_j(i)$ it is possible to derive the following recursive equation for X_k :

$$(4.22) \quad X_{k+4} = \frac{f(X_{k+3}, X_{k+2}, X_{k+1})}{g(X_{k+3}, X_{k+2}, X_{k+1})},$$

where

$$\begin{aligned} f(X_{k+3}, X_{k+2}, X_{k+1}) = & \phi_k(1)X_{k+3} + \phi_k(2)X_{k+2} + \phi_k(3)X_{k+1} + \phi_k(4)X_{k+3}X_{k+2} \\ & + \phi_k(5)X_{k+2}X_{k+1} + \phi_k(6)X_{k+3}X_{k+2}X_{k+1} + \phi_k(7) \end{aligned}$$

and

$$\begin{aligned} g(X_{k+3}, X_{k+2}, X_{k+1}) = & \phi_k(8)X_{k+3} + \phi_k(9)X_{k+2} + \phi_k(10)X_{k+1} + \phi_k(11)X_{k+3}X_{k+2} \\ & + \phi_k(12)X_{k+3}X_{k+1} + \phi_k(13)X_{k+2}X_{k+1} + \phi_k(14)X_{k+3}X_{k+2}X_{k+1} + \phi_k(15). \end{aligned}$$

In the above equation, $\phi_k(i)i = 1, \dots, 15$ are nonlinear functions of the plant parameters and the parameters to be assigned at times $k+1, k+2, k+3$.

A complete analysis of recursions of the type (4.22) is a subject of future research. We would analyze (4.22) under certain special cases. If we denote X_k by y_k/ζ_k , we can rewrite the recursion (4.22) as follows.

$$\begin{aligned} y_{k+4} = & \phi_k(1)y_{k+3}\zeta_{k+2}\zeta_{k+1} + \phi_k(2)\zeta_{k+3}y_{k+2}\zeta_{k+1} \\ & + \phi_k(3)\zeta_{k+3}\zeta_{k+2}y_{k+1} + \phi_k(4)y_{k+3}y_{k+2}\zeta_{k+1} \\ & + \phi_k(5)\zeta_{k+3}y_{k+2}y_{k+1} + \phi_k(6)y_{k+3}y_{k+2}y_{k+1} + \phi_k(7)\zeta_{k+3}\zeta_{k+2}\zeta_{k+1} \cdot \\ \zeta_{k+4} = & \phi_k(8)y_{k+3}\zeta_{k+2}\zeta_{k+1} + \phi_k(9)\zeta_{k+3}y_{k+2}\zeta_{k+1} + \phi_k(10)\zeta_{k+3}\zeta_{k+2}y_{k+1} \\ & + \phi_k(11)y_{k+3}y_{k+2}\zeta_{k+1} + \phi_k(12)\zeta_{k+3}y_{k+2}y_{k+1} \\ & + \phi_k(13)y_{k+3}\zeta_{k+2}y_{k+1} + \phi_k(14)\zeta_{k+3}\zeta_{k+2}\zeta_{k+1} + \phi_k(15)y_{k+3}y_{k+2}y_{k+1}. \end{aligned}$$

If we assume without any loss of generality that $y_k^2 + \zeta_k^2 = 1$, we can reparameterize $y_k = \cos \theta_k, \zeta_k = \sin \theta_k$. Furthermore, if we choose

$$(4.23) \quad \begin{aligned} \phi_k(1) = \phi_k(2) = \phi_k(3) = -1, \phi_k(6) = \phi_k(4) = \phi_k(5) = \phi_k(7) = 0, \\ \phi_k(8) = \phi_k(9) = \phi_k(10) = \phi_k(15) = 0, \\ \phi_k(11) = \phi_k(12) = \phi_k(13) = 1, \phi_k(14) = -1, \end{aligned}$$

we have

$$\begin{aligned}\cos \theta_{k+4} &= \cos(\theta_{k+3} + \theta_{k+2} + \theta_{k+1}), \\ \sin \theta_{k+4} &= \sin(\theta_{k+3} + \theta_{k+2} + \theta_{k+1}).\end{aligned}$$

Thus, for the special choices of $\phi_k(\cdot)$ given by (4.23), the recursion (4.22) reduces to

$$(4.24) \quad \theta_{k+4} = \theta_{k+3} + \theta_{k+2} + \theta_{k+1}.$$

We now claim that (4.24) describes an Anosov flow [32] on T^3 , the three-dimensional torus. Consider the system

$$(4.25) \quad \begin{bmatrix} \alpha_{k+1} \\ \beta_{k+1} \\ \nu_{k+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_k \\ \beta_k \\ \nu_k \end{bmatrix}$$

on \mathbb{R}^3 , where $\alpha_0 = \theta_0$, $\beta_0 = \theta_1$, and $\nu_0 = \theta_2$. It follows that $\nu_k = \theta_{k+2}$. Let the 3×3 matrix in (4.25) be denoted by A . Since all entries of A are integers, $\det A = 1$ and A is hyperbolic, it follows that the map induced on T^3 by A is a hyperbolic toral automorphism, which we denote by L_A . It follows from [32, Thm. 4.8] that periodic points of L_A are dense in T^3 , L_A is topologically transitive, and L_A has sensitive dependence on initial conditions. Thus, the hyperbolic toral automorphism is chaotic on all of T^3 (see [32, p. 197]).

What we conclude, therefore, is that for choices of $\phi_k(\cdot)$ given by (4.23), the time-varying coefficient assignment problem is chaotic. Thus, if we use adaptive coefficient assignment as a strategy for compensation, we must carefully avoid chaotic dynamics.

5. Conclusion. In this paper, we have given conditions under which an input–output map for a time-varying system admits left and/or right matrix fraction representations. Using these representations, we have described procedures for the simultaneous coefficient assignment of a family of time-varying systems using nonrecursive algebraic techniques. It is important to note that these techniques are generalizations of well-known design methodologies in time-invariant system theory [1]–[3] to the time-varying case. When the number of input and output channels is such that the conditions of the nonrecursive coefficient assignment scheme is not satisfied, recursive procedures to coefficient assign time-varying systems are given in the form of a recursive algorithm. These recursive procedures have no analogues in the time-invariant case, and thus, represent a new design procedure. For certain special cases, solution of the proposed recursive algorithm is shown to be chaotic. This fact indicates that further work needs to be done and a complete analysis of the algorithms given in §4 is a subject of future research.

6. Appendix I. The purpose of this appendix is to prove Theorem 2.13. Let us consider a right factorization of lag $\bar{\ell}$ given by (2.24). Let us also consider a left factorization of the form (2.12) of lag ℓ . The two representations have the same input–output properties provided:

$$(6.1) \quad \left[\sum_{i=0}^{\bar{\ell}} \bar{N}_k(i) z^{-i} \right] \circ \left[I + \sum_{i=1}^{\bar{\ell}} \bar{D}_k(i) z^{-i} \right]^{-1} = \left[I + \sum_{i=1}^{\ell} D_k(i) z^{-i} \right]^{-1} \circ \left[\sum_{i=0}^{\ell} N_k(i) z^{-i} \right]$$

or, equivalently,

$$\left[I + \sum_{i=1}^{\ell} D_k(i) z^{-i} \right] \circ \left[\sum_{i=0}^{\bar{\ell}} \bar{N}_k(i) z^{-i} \right] - \left[\sum_{i=0}^{\ell} N_k(i) z^{-i} \right] f \circ \left[I + \sum_{i=1}^{\bar{\ell}} \bar{D}_k(i) z^{-i} \right] = 0. \quad (6.2)$$

Expanding (6.2) and equating like powers of z^{-1} , we obtain

$$N_k(0) = \bar{N}_k(0) \text{ for } k = 0, 1, 2, \dots$$

and

$$\nu_k M_k = \phi_k \text{ for } k = 0, 1, 2, \dots, \quad (6.3)$$

where

$$\begin{aligned} \nu_k &= [D_k(1), D_k(2), \dots, D_k(\ell), N_k(1), N_k(2), \dots, N_k(\ell)], \\ \phi_k &= [\bar{N}_k(0)\bar{D}_k(1) - \bar{N}_k(1), \dots, \bar{N}_k(0)\bar{D}_k(\bar{\ell}) - \bar{N}_k(\bar{\ell}); 0, \dots, 0], \end{aligned}$$

$$M(k) =$$

$$\begin{bmatrix} \bar{N}_{k-1}(0) & \bar{N}_{k-1}(1) & \bar{N}_{k-1}(2) & \cdots & \bar{N}_{k-1}(\bar{\ell}) & O & O & O \\ O & \bar{N}_{k-2}(0) & \bar{N}_{k-2}(1) & \cdots & \cdots & \bar{N}_{k-2}(\bar{\ell}) & O & O \\ & & \vdots & & & & & \\ O & O & \bar{N}_{k-\ell}(0) & \cdots & \cdots & & & \bar{N}_{k-\ell}(\bar{\ell}) \\ -I & -\bar{D}_{k-1}(1) & -\bar{D}_{k-1}(2) & \cdots & -\bar{D}_{k-1}(\bar{\ell}) & O & O & O \\ O & -I & -\bar{D}_{k-2}(1) & \cdots & \cdots & -\bar{D}_{k-2}(\bar{\ell}) & O & O \\ & & \vdots & & & & & \\ O & O & \cdots & -I & -\bar{D}_{k-\ell}(1) & \cdots & & -\bar{D}_{k-\ell}(\bar{\ell}) \end{bmatrix}. \quad (6.4)$$

As ν_k is a $p \times \ell(p+m)$ matrix, ϕ_k is a $p \times (\ell + \bar{\ell})m$ matrix and M_k is a $\ell(p+m) \times (\ell + \bar{\ell})m$ matrix; it follows that a sufficient condition for (6.3) to have a solution ν_k is that M_k is of full row rank and $\ell(p+m) \geq (\ell + \bar{\ell})m$, i.e., if $\ell p \geq \bar{\ell}m$.

It is not too hard to check (see [27] for details) that the condition that M_k is not of full row rank is given by a proper algebraic set in \mathcal{P} (in the topology described in §2 (25)). In fact, for $k = \ell + \tau$, the condition that M_k is not of full row rank is obtained as proper algebraic set in the restriction of \mathcal{P} to

$$\prod_{j=\tau}^{\ell-1+\tau} \mathbb{R}^N$$

for $\tau = 0, 1, 2, \dots$. Thus, there is a countable intersection of open and dense set in \mathcal{P} for which M_k is of full row rank for all $k \geq \ell$.

Remark. In general, it is not entirely obvious why the associated algebraic sets in \mathcal{P} that makes M_k singular is proper. The proof consists of picking a minor of M_k with nonidentically vanishing determinant. The details being technical are relegated to [27].

REFERENCES

- [1] B. K. GHOSH, *An approach to simultaneous system design, part I: semialgebraic geometric methods*, SIAM J. Control Optim., 24 (1986), pp. 480–496.
- [2] ———, *An approach to simultaneous system design, part II: nonswitching gain and dynamic feedback compensation by algebraic geometric methods*, SIAM J. Control Optim., 26 (1988), pp. 919–963.
- [3] ———, *Transcendental and interpolation methods in simultaneous stabilization and simultaneous partial pole placement problems*, SIAM J. Control Optim., 24 (1986), pp. 1091–1109.
- [4] B. K. GHOSH AND C. I. BYRNES, *Simultaneous stabilization and simultaneous pole placement by nonswitching dynamic compensation*, IEEE TRans. Automat. Control, 28 (1983), pp. 735–741.
- [5] M. VIDYASAGAR AND N. VISWANDADHAM, *Algebraic design techniques for reliable stabilization*, IEEE Trans. Automat. Control, AC-27 (Oct. 1983), pp. 1085–1095.
- [6] R. SAEKS AND J. MURRAY, *Fractional representation, algebraic geometry and the simultaneous stabilization problem*, IEEE Trans. Automat. Control, AC-27 (Aug. 1982), pp. 895–903.
- [7] K. D. MINTO, *Design of reliable control systems: theory and computation*, Ph.D. dissertation, Dept. of Electrical Engineering, University of Waterloo, Waterloo, Canada, 1985.
- [8] M. Y. WU AND A. SHERIF, *On explicit solution, stability and reduction of a class of linearly time-varying discrete-time systems*, Internat. J. Control, 25, 2 (1977), pp. 303–310.
- [9] M. Y. WU, *A note on stability of linear time-varying systems*, IEEE Trans. Automat. Control, AC-19 (1974), p. 162.
- [10] C. A. DESOER, *Slowly varying discrete system $x_{i+1} = A_i x_i$* , Electron. Letters, 6, 11 (1970), pp. 339–340.
- [11] P. R. BOUTHELLIER AND B. K. GHOSH, *A Stability Theory of Linear Time-Varying Systems*, Proc. 25th Annual Allerton Conference on Communication, Control and Computing (1987), pp. 1172–1180.
- [12] ———, *A New Class of Stable Time-Varying Systems and the Coefficient Assignment Problem*, Proc. 28th IEEE CDC, pp. 2343–2347, 1989.
- [13] B. AULBACH, *Continuous and Discrete Dynamics Near Manifolds of Equilibria*, Springer-Verlag, Berlin, 1984.
- [14] K. R. POOLLA, *Linear time-varying systems: representations and control via transfer function matrices*, Ph.D. dissertation, University of Florida, Gainesville, 1984.
- [15] E. W. KAMEN, P. P. KHARGONEKAR, AND K. R. POOLLA, *A transfer-function approach to linear time-varying systems*, SIAM J. Control Optim., 24 (1986), pp. 550–565.
- [16] S. K. HWANG, *An augmentation approach and theory of the resultant for linear time-varying systems*, Ph.D. dissertation, University of Florida, Gainesville, 1986.
- [17] P. P. KHARGONEKAR AND K. R. POOLLA, *Polynomial matrix-fraction representations for linear time-varying systems*, Linear Algebra Appl., 80 (1986), pp. 1–17.
- [18] K. S. TSAKALIS AND P. A. IOANNOU, *A new indirect adaptive control scheme for time-varying plants*, Proc. 27th IEEE CDC, 1988, pp. 2419–2424.
- [19] ———, *Adaptive control of linear time-varying plants*, Automat., 23 (1987), pp. 459–468.
- [20] P. K. STEVENS, *Algebra-geometric methods for linear multivariable feedback systems*, Ph.D. dissertation, Harvard University, Cambridge, MA, 1982.
- [21] T. DJAFERIS, *Robust observers and regulation for systems with parameters*, Proc. 23rd IEEE CDC, 1984, pp. 1234–1239.
- [22] F. M. BRASCH AND J. B. PEARSON, *Pole placement using dynamic compensators*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 34–43.
- [23] P. P. KHARGONEKAR, A. M. PASCOAL, AND R. RAVI, *Strong, simultaneous, and reliable stabilization of finite-dimensional linear time-varying plants*, IEEE Trans. Automat. Control, 33, 12 (Dec. 1988), pp. 1158–1163.
- [24] M. A. ROTEA AND P. P. KHARGONEKAR, *Stabilizability of linear time-varying and uncertain linear systems*, IEEE Trans. Automat. Control, 33 (Sept. 1988), pp. 884–887.
- [25] M. VIDYASAGER, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
- [26] T. KAILATH, *Linear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1980.
- [27] P. R. BOUTHELLIER, *Analysis and design of discrete-time, linear time-varying systems*, D.Sc. dissertation, Washington University, St. Louis, MO, 1990.
- [28] S. WILLARD, *General Topology*, Addison Wesley, Reading, MA, 1970.
- [29] B. D. O. ANDERSON AND C. I. BYRNES, *Output feedback and generic stabilizability*, SIAM J. Control Optim., 22 (1984), pp. 362–380.
- [30] R. W. BROCKETT, *Finite dimensional linear systems*, John Wiley, New York, 1969.
- [31] P. R. BOUTHELLIER AND B. K. GHOSH, *Robust stabilization of discrete-time, single-input single-output, linear time-varying systems*, Proc. 27th IEEE CDC, pp. 39–44, 1988.
- [32] R. L. DEVANEY, *An Introduction to Chaotic Dynamical Systems*, Addison Wesley, Reading, MA, 1987.

ON THE NECESSARY CONDITIONS OF OPTIMAL CONTROLS FOR STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS*

XUN YU ZHOU†

Abstract. This paper concerns optimal control of systems governed by stochastic partial differential equations in which drift and diffusion terms are second- and first-order differential operators, respectively. Necessary conditions for an optimal control are derived for both nondegenerate and degenerate systems, and all the coefficients appearing in the equations are allowed to depend on the control variables. Furthermore, the results obtained in the paper can also be used to derive necessary conditions of optimality for partially observed diffusions with correlation between the signals and the observation noises.

Key words. stochastic partial differential equations (SPDEs), optimal control, adjoint equations, necessary conditions of optimality, partially observed diffusions.

AMS subject classifications. 60H15, 93E20

1. Introduction. Let us consider a stochastic control problem, in which the state equation is a linear stochastic partial differential equation (SPDE):

$$(1.1) \quad \begin{aligned} dq(t, x) &= [\partial_i(a^{ij}(t, x, U(t))\partial_j q(t, x)) + b^i(t, x, U(t))\partial_i q(t, x) + c(t, x, U(t))q(t, x) \\ &\quad + f(t, x, U(t))]dt + [\sigma^{ik}(t, x, U(t))\partial_i q(t, x) + h^k(t, x, U(t))q(t, x) \\ &\quad + g^k(t, x, U(t))]dW_k(t), \quad x \in R^d, \quad t \in [0, 1], \\ q(0, x) &= q_0(x), \quad x \in R^d, \end{aligned}$$

where $W := (W_1, W_2, \dots, W_{d'})$ is a d' -dimensional standard Brownian motion with $W(0) = 0$, $\{U(t) : 0 \leq t \leq 1\}$ is an admissible control (the precise definition will be given later) and $\partial_i := \partial/\partial x_i$, $i = 1, 2, \dots, d$. Throughout the paper, the conventional repeated indices for summation are used.

The optimal control problem is to minimize a given cost functional over the set of admissible controls.

The purpose of this paper is to study necessary conditions of an optimal control for the controlled system (1.1). It is well known that the so-called adjoint equations play a key role in dealing with the problem. Adjoint equations are, in general, *backward* equations with given terminal states. In stochastic problems, adjoint equations cannot be obtained simply by inverting the time, because the adaptiveness must be considered. For stochastic differential equations (SDEs), Bismut [5] introduced an adjoint equation with an additional martingale term. His method is based on the invertibility of certain fundamental matrices in finite dimensions and cannot be carried over to SPDEs whose state spaces are *infinite* dimensions. Using a finite-dimensional approximation method, Bensoussan [3] derived an adjoint equation of a nondegenerate SPDE (in a form more abstract than (1.1)) with its diffusion being a bounded operator. In the present paper, we solve the adjoint equation of (1.1) with $\sigma^{ik} \neq 0$ (i.e., the diffusion operator is unbounded), which is important in both theory and application. To handle the problem, the basic approach we employ is still the finite-dimensional approximation. It should be noted that, while this is a very natural approach to infinite-dimensional problems, the difficulty is how to obtain certain *compactness* of the approximate solutions, which may vary from case to case. For explicit equations like

* Received by the editors May 28, 1991; accepted for publication (in revised form) May 1, 1992. This research was supported partially by the Monbusho Scholarship of Japanese Government and the National Natural Science Foundation of China.

† Department of Mathematics, Fudan University, Shanghai, China. This research was completed during a visit by the author to Department of Mathematics, Faculty of Science, Kobe University, Kobe 657, Japan.

(1.1), some delicate estimates of differential operators, originally due to Krylov and Rozovskii [8]–[10], will be used in this paper to show the compactness. Therefore, an adjoint equation of (1.1) as well as existence and uniqueness of its solutions will be derived.

Having obtained the adjoint equation, we derive necessary conditions of optimality for system (1.1) in which *all* the coefficients are allowed to depend on the control variable. Furthermore, the results obtained can be applied directly to a general model of partially observed diffusions with *correlation* between the signals and the observation noises. Therefore the existing results of Bensoussan [3], [4], Haussmann [7], and Baras, Elliott, and Kohlmann [1] are improved and extended considerably.

It should be mentioned that, in addition to the finite-dimensional approximation approach, there is a “time change” technique, which Bensoussan [4] used to solve the adjoint equation and study the necessity of optimality for the SPDE (1.1) with $\sigma^{ik} = 0$. Its main idea is to turn (1.1) into a P -almost surely *deterministic* PDE, basing on a transformation $e^{M^k W_k(t)}$, where M^k represents the diffusion operator (see (2.2), below). When $\sigma^{ik} \neq 0$, transformations of the same kind are also available, provided that the Brownian motion involved is one-dimensional; see [6], [15]. However, this method fails to work in general when $\sigma^{ik} \neq 0$ and the Brownian motion is multidimensional; refer to [15, §5.1] for a detailed discussion on this point.

The paper is organized as follows. In §2 we formulate our problem and introduce some basic notation and assumptions. In §3 we state a few fundamental results of SPDEs in the form convenient for us to use in the paper. Sections 4 and 5 are the main part of the paper. In §4 we derive an adjoint equation and prove existence and uniqueness of its solutions, and in §5 we investigate necessary conditions of optimality for system (1.1). The results in these two sections are valid, assuming that the system is nondegenerate (i.e., $S := (a^{ij} - \frac{1}{2} \sum_{k=1}^{d'} \sigma^{ik} \sigma^{jk})$ is uniformly positive definite). In §6 we discuss the degenerate case (i.e., S is nonnegative definite) and apply the main results to partially observed diffusions. Finally, §7 concludes the paper.

2. Problem formulation. Let Γ be a Borel set in some Euclidean space R^M . We define a family of second-order differential operators $\{A(t, u) : t \in [0, 1], u \in \Gamma\}$ and a family of first-order differential operators $\{M^k(t, u) : t \in [0, 1], u \in \Gamma, k = 1, 2, \dots, d'\}$ by

$$(2.1) \quad A(t, u)\phi(x) := \partial_i(a^{ij}(t, x, u)\partial_j\phi(x)) + b^i(t, x, u)\partial_i\phi(x) + c(t, x, u)\phi(x)$$

and

$$(2.2) \quad M^k(t, u)\phi(x) := \sigma^{ik}(t, x, u)\partial_i\phi(x) + h^k(t, x, u)\phi(x) \quad \text{for } x \in R^d, \phi \in C_0^\infty(R^d),$$

where $a^{ij}, b^i, c, \sigma^{ik}$, and h^k are given real-valued functions, $i, j = 1, 2, \dots, d$ and $k = 1, 2, \dots, d'$.

We also consider the formal adjoints of the operators $A(t, u)$ and $M^k(t, u)$

$$(2.3) \quad \begin{aligned} A^*(t, u)\phi(x) := & \partial_i(a^{ij}(t, x, u)\partial_j\phi(x)) - b^i(t, x, u)\partial_i\phi(x) \\ & + [c(t, x, u) - \partial_i b^i(t, x, u)]\phi(x), \end{aligned}$$

$$(2.4) \quad M^{k*}(t, u)\phi(x) := -\sigma^{ik}(t, x, u)\partial_i\phi(x) + [h^k(t, x, u) - \partial_i \sigma^{ik}(t, x, u)]\phi(x).$$

Let us now recall the definition of the Sobolev spaces. For $m = 0, 1, 2, \dots$, define $H^m := \{\phi : D^\alpha \phi \in L^2(R^d) \text{ for any } \alpha := (\alpha_1, \dots, \alpha_d) \text{ with } |\alpha| := |\alpha_1| + \dots + |\alpha_d| \leq m\}$,

with the norm

$$\|\phi\|_m := \left\{ \sum_{|\alpha| \leq m} \int_{R^d} |D^\alpha \phi(x)|^2 dx \right\}^{1/2}.$$

For $m = -1, -2, \dots$, define $H^m := (H^{-m})^*$. The Hilbert space H^m for any integer m is called a Sobolev space. For any integer m , let us consider the Gelfand triple $H^{m+1} \hookrightarrow H^m \hookrightarrow H^{m-1}$. We denote by $\langle \cdot, \cdot \rangle_m$ the duality pairing between H^{m-1} and H^{m+1} , and by $(\cdot, \cdot)_m$ the inner product in H^m .

For any second-order differential operator L that has the same form as (2.1), if we write $\langle L\phi, \psi \rangle_m$, then L is understood to be an operator from H^{m+1} to H^{m-1} by formally using Green's formula. For example, for the operator $A(t, u)$ defined by (2.1), we have

$$(2.5) \quad \begin{aligned} \langle A(t, u)\phi, \psi \rangle_m &:= -(a^{ij}(t, \cdot, u)\partial_j \phi, \partial_i \psi)_m + (b^i(t, \cdot, u)\partial_i \phi, \psi)_m \\ &\quad + (c(t, \cdot, u)\phi, \psi)_m \quad \text{for } \phi, \psi \in H^{m+1}. \end{aligned}$$

Remark 2.1. It is clear that $\langle A(t, u)\phi, \psi \rangle_0 = \langle \phi, A^*(t, u)\psi \rangle_0$ and $(M(t, u)\phi, \psi)_0 = (\phi, M^*(t, u)\psi)_0$ hold for $\phi, \psi \in H^1$. However, neither $\langle A(t, u)\phi, \psi \rangle_m = \langle \phi, A^*(t, u)\psi \rangle_m$ nor $(M(t, u)\phi, \psi)_m = (\phi, M^*(t, u)\psi)_m$ holds when $m \geq 1$.

For $\alpha, \beta \in (-\infty, +\infty)$ with $\alpha < \beta$, we are given a filtered probability space $(\Omega, \mathcal{F}, P, \mathcal{F}_t : \alpha \leq t \leq \beta)$ and a Hilbert space X . For $p \in [1, +\infty]$, define $L_{\mathcal{F}}^p(\alpha, \beta, X) := \{\phi : \phi \text{ is an } X\text{-valued, } \mathcal{F}_t\text{-adapted process on } [\alpha, \beta], \text{ and } \phi \in L^p([\alpha, \beta] \times \Omega; X)\}$. We identify ϕ and ϕ' in $L_{\mathcal{F}}^p(\alpha, \beta; X)$ whenever $D \int_{\alpha}^{\beta} \|\phi(t) - \phi'(t)\|_X^p dt = 0$.

Now we recall the definition of admissible controls. By the set U_{ad} of admissible controls, we mean the collection of (i) standard probability spaces (Ω, \mathcal{F}, P) and d' -dimensional Brownian motions $\{W(t) : 0 \leq t \leq 1\}$ with $W(0) = 0$; (ii) Γ -valued, \mathcal{F}_t -adapted measurable processes $\{U(t) : 0 \leq t \leq 1\}$, where $\mathcal{F}_t := \sigma\{W(s) : 0 \leq s \leq t\}$. We denote $(\Omega, \mathcal{F}, P, W, U) \in U_{\text{ad}}$, but on occasion we will only write $U \in U_{\text{ad}}$ if no ambiguity arises.

Given $(\Omega, \mathcal{F}, P, W, U) \in U_{\text{ad}}$, we rewrite (1.1) in the following form, omitting the variable x :

$$(2.6) \quad \begin{aligned} dq(t) &= [A(t, U(t))q(t) + f(t, U(t))]dt \\ &\quad + [M^k(t, U(t))q(t) + g^k(t, U(t))]dW_k(t), \quad t \in [0, 1], \\ q(0) &= q_0. \end{aligned}$$

A process $q = q^U \in L_{\mathcal{F}}^2(0, 1; H^1)$ is called a solution of (2.6) or a response for the control U if, for each $\eta \in C_0^\infty(R^d)$ and almost all $(t, \omega) \in [0, 1] \times \Omega$,

$$(2.7) \quad \begin{aligned} (q(t), \eta)_0 &= (q_0, \eta)_0 + \int_0^t \langle A(s, U(s))q(s) + f(s, U(s)), \eta \rangle_0 ds \\ &\quad + \int_0^t (M^k(s, U(s))q(s) + g^k(s, U(s)), \eta)_0 dW_k(s). \end{aligned}$$

The optimal control problem is to choose $(\Omega, \mathcal{F}, P, W, U) \in U_{\text{ad}}$ to minimize the following cost functional:

$$(2.8) \quad J(U) := E \left[\int_0^1 \langle F(t, U(t)), q^U(t) \rangle_0 dt + (G, q^U(1))_0 \right],$$

where $F : [0, 1] \times \Gamma \rightarrow H^{-1}$ and $G \in H^0$ are given.

Remark 2.2. The cost functional (2.8) includes the following one as a special case:

$$J(U) := E \left\{ \int_0^1 [(F^0(t, U(t)), q^U(t))_0 + (F^i(t, U(t)), \partial_i q^U(t))_0] dt + (G, q^U(1))_0 \right\}.$$

Let us fix an integer $m \geq 0$ and two positive constants K and δ . We introduce the following conditions:

(A1) $_m$ $a^{ij}, b^i, c, \sigma^{ik}, h^k : [0, 1] \times R^d \times \Gamma \rightarrow R^1$ are measurable in (t, x, u) and continuous in u . Furthermore, the functions $a^{ij}, b^i, c, \sigma^{ik}, h^k, \partial_j \sigma^{ik}$, and $\partial_j h^k$ and their derivatives in x up to the order $\max(2, m)$ do not exceed K in absolute value;

(A2) $a^{ij} = a^{ji}, i, j = 1, 2, \dots, d$, and the matrix $S := (a^{ij} - \frac{1}{2} \sum_{k=1}^{d'} \sigma^{ik} \sigma^{jk}) \geq 0$ for all (t, x, u) ;

(A2)' $a^{ij} = a^{ji}, i, j = 1, 2, \dots, d$, and the matrix S is uniformly positive definite:

$$\xi^T S \xi \geq \delta |\xi|^2 \quad \text{for any } (t, x, u) \text{ and any } \xi \in R^d;$$

(A3) $F(t, u) \in H^{-1}, G \in H^0$ and $\|F(t, u)\|_{-1} + \|G\|_0 \leq K$;

(A4) $f, g^k : [0, 1] \times R^d \times \Gamma \rightarrow R^1$ are measurable in (t, x, u) and continuous in $u, k = 1, 2, \dots, d'$. Furthermore, $f(t, \cdot, u) \in H^1, g^k(t, \cdot, u) \in H^2$, and

$$|f(t, x, u)| + |g^k(t, x, u)| + \|f(t, \cdot, u)\|_1 + \|g^k(t, \cdot, u)\|_2 \leq K, \quad k = 1, 2, \dots, d';$$

(A5) $q_0 \in H^1$.

3. Fundamental theory of SPDEs. In this section, we recall some fundamental facts about SPDEs, which are originally due to Krylov and Rozovskii [8]–[10]. We state the results and give some variants that are convenient to our later discussion.

PROPOSITION 3.1. (Krylov and Rozovskii [9]). *Let \tilde{A} be any second-order differential operator having the same form as (2.1), and let \tilde{M}^k be any first-order differential operator having the same form as (2.2). Assume that the coefficients of \tilde{A} and \tilde{M}^k satisfy (A1) $_m$ and (A2). Then there is a constant N_1 that depends only on K and m such that*

$$2[\langle \tilde{A}\phi, \phi \rangle_{\tilde{m}} + (\tilde{f}, \phi)_{\tilde{m}}] + \sum_{k=1}^{d'} \|\tilde{M}^k \phi + \tilde{g}^k\|_{\tilde{m}}^2 \leq N_1 \left(\|\phi\|_{\tilde{m}}^2 + \|\tilde{f}\|_{\tilde{m}}^2 + \sum_{k=1}^{d'} \|\tilde{g}^k\|_{\tilde{m}+1}^2 \right) \\ \text{for any } \phi \in H^{\tilde{m}+1}, \tilde{f} \in H^{\tilde{m}}, \text{ and } \tilde{g}^k \in H^{\tilde{m}+1}, \tilde{m} = 0, \dots, m. \quad (3.1)$$

Remark 3.1. Estimate (3.1) is slightly different in form from the original result of Krylov and Rozovskii [9]. We may consult Nagase and Nisio [12, §3 and Appendix] for an explicit proof of (3.1).

COROLLARY 3.1. *In Proposition 3.1., let us assume, in addition, that the operator \tilde{M}^k is of order zero. Then there is a constant N_2 that depends only on K and m such that*

$$2[\langle \tilde{A}\phi, \phi \rangle_{\tilde{m}} + (\tilde{f}, \phi)_{\tilde{m}}] + \sum_{k=1}^{d'} \|\tilde{M}^k \phi + \tilde{g}^k\|_{\tilde{m}}^2 \leq N_2 \left(\|\phi\|_{\tilde{m}}^2 + \|\tilde{f}\|_{\tilde{m}}^2 + \sum_{k=1}^{d'} \|\tilde{g}^k\|_{\tilde{m}}^2 \right) \\ \text{for any } \phi \in H^{\tilde{m}+1}, \text{ and } \tilde{f}, \tilde{g}^k \in H^{\tilde{m}}, \tilde{m} = 0, 1, \dots, m. \quad (3.2)$$

Proof. From (3.1), it follows that

$$2[\langle \tilde{A}\phi, \phi \rangle_{\tilde{m}} + (\tilde{f}, \phi)_{\tilde{m}}] \leq N_1 (\|\phi\|_{\tilde{m}}^2 + \|\tilde{f}\|_{\tilde{m}}^2). \quad (3.3)$$

By the hypotheses of the corollary, \tilde{M}^k becomes a bounded operator on $H^{\bar{m}}$. Hence

$$(3.4) \quad \|\tilde{M}^k \phi + \tilde{g}^k\|_{\bar{m}}^2 \leq 2(K^2 \|\phi\|_{\bar{m}}^2 + \|\tilde{g}^k\|_{\bar{m}}^2).$$

Therefore, estimate (3.2) follows from (3.3) and (3.4). \square

COROLLARY 3.2. *In Proposition 3.1, let us assume, in addition, that (A2)' is satisfied. Then there is a constant N_3 that depends only on K, m , and δ such that*

$$(3.5) \quad \begin{aligned} & 2[\langle \tilde{A}\phi, \phi \rangle_{\bar{m}} + \langle \tilde{f}, \phi \rangle_{\bar{m}}] + \sum_{k=1}^{d'} \|\tilde{M}^k \phi + \tilde{g}^k\|_{\bar{m}}^2 \\ & \leq -\delta \|\phi\|_{\bar{m}+1}^2 + N_3 \left(\|\phi\|_{\bar{m}}^2 + \|\tilde{f}\|_{\bar{m}-1}^2 + \sum_{k=1}^{d'} \|\tilde{g}^k\|_{\bar{m}}^2 \right) \\ & \quad \text{for } \phi \in H^{\bar{m}+1}, \tilde{f} \in H^{\bar{m}-1} \text{ and } \tilde{g}^k \in H^{\bar{m}}, \bar{m} = 0, 1, \dots, m. \end{aligned}$$

Proof. The result can be easily derived by applying Proposition 3.1 to the operator $A - \delta \Delta$, where Δ is the Laplacian. \square

Remark 3.2. When $m = 0$ (3.5) holds even if all the coefficients of \tilde{A} and \tilde{M}^k are only bounded measurable in x -variable; see Pardoux [13].

Let $\tilde{F}, \tilde{G}^k : [0, 1] \times R^d \times \Omega \rightarrow R^1, \xi_0 : R^d \times \Omega \rightarrow R^1$ be given. We now consider the following SPDE on an interval $[\alpha, \beta] \subset [0, 1]$:

$$(3.6) \quad \begin{aligned} d\xi(t) &= [A(t, U(t))\xi(t) + \tilde{F}(t)]dt \\ &\quad + [M^k(t, U(t))\xi(t) + \tilde{G}^k(t)]dW_k(t), \quad t \in [\alpha, \beta], \\ \xi(\alpha) &= \xi_0, \end{aligned}$$

PROPOSITION 3.2 (Krylov and Rozovskii [9], [10]). *Assume that (A1) $_m$ and (A2) ($m \geq 1$) are satisfied and that $\tilde{F} \in L^2_{\mathcal{F}}(0, 1; H^m), \tilde{G}^k \in L^2_{\mathcal{F}}(0, 1; H^{m+1})$, and $\xi_0 \in L^2(\Omega, \mathcal{F}_\alpha, P; H^m)$. Then (3.6) has a unique solution $\xi \in L^2_{\mathcal{F}}(\alpha, \beta; H^m) \cap L^2(\Omega; C(\alpha, \beta; H^{m-1}))$, and there is a constant N_4 that depends only on K and m such that*

$$(3.7) \quad E \left(\sup_{\alpha \leq t \leq \beta} \|\xi(t)\|_{\bar{m}}^2 \right) \leq N_4 E \left\{ \|\xi_0\|_{\bar{m}}^2 + \int_{\alpha}^{\beta} \left[\|\tilde{F}(t)\|_{\bar{m}}^2 + \sum_{k=1}^{d'} \|\tilde{G}^k(t)\|_{\bar{m}+1}^2 \right] dt \right\},$$

$\bar{m} = 0, 1, \dots, m.$

COROLLARY 3.3. *Assume that (A1) $_m$ and (A2) ($m \geq 1$) are satisfied with $\sigma^{ik} = 0$ and that $\tilde{F}, \tilde{G}^k \in L^2_{\mathcal{F}}(0, 1; H^m)$ and $\xi_0 \in L^2(\Omega, \mathcal{F}_\alpha, P; H^m)$. Then (3.6) has a unique solution $\xi \in L^2_{\mathcal{F}}(\alpha, \beta; H^m) \cap L^2(\Omega; C(\alpha, \beta; H^{m-1}))$, and estimate (3.7) can be strengthened to*

$$(3.8) \quad E \left(\sup_{\alpha \leq t \leq \beta} \|\xi(t)\|_{\bar{m}}^2 \right) \leq N_5 E \left\{ \|\xi_0\|_{\bar{m}}^2 + \int_{\alpha}^{\beta} \left[\|\tilde{F}(t)\|_{\bar{m}}^2 + \sum_{k=1}^{d'} \|\tilde{G}^k(t)\|_{\bar{m}}^2 \right] dt \right\},$$

$\bar{m} = 0, 1, \dots, m.$

Proof. Note that estimate (3.2) is available as $\sigma^{ik} = 0$; hence the result can be proved in a way analogous to that in Krylov and Rozovskii [9]. \square

PROPOSITION (3.3) (Krylov and Rozovskii [8]). Assume that $(A1)_m$ and $(A2)' (m \geq 0)$ are satisfied and that $\tilde{F} \in L^2_{\mathcal{F}}(0, 1; H^{m-1})$, $\tilde{G}^k \in L^2_{\mathcal{F}}(0, 1; H^m)$, and $\xi_0 \in L^2(\Omega, \mathcal{F}_\alpha, P; H^m)$. Then (3.6) has a unique solution $\xi \in L^2_{\mathcal{F}}(\alpha, \beta; H^{m+1}) \cap L^2(\Omega; C(\alpha, \beta; H^m))$, and there is a constant N_6 that depends only on K, m , and δ such that

$$E \left(\sup_{\alpha \leq t \leq \beta} \|\xi(t)\|_{\tilde{m}}^2 \right) \leq N_6 E \left\{ \|\xi_0\|_{\tilde{m}}^2 + \int_{\alpha}^{\beta} \left[\|\tilde{F}(t)\|_{\tilde{m}-1}^2 + \sum_{k=1}^{d'} \|\tilde{G}^k(t)\|_{\tilde{m}}^2 \right] dt \right\},$$

$\tilde{m} = 0, 1, \dots, m.$

(3.9)

4. Adjoint equation. In this section, we derive an adjoint equation and study existence and uniqueness of its solutions. We fix an admissible control $(\Omega, \mathcal{F}, P, W, U)$ throughout this section.

THEOREM 4.1. Assume that $(A1)_0$, $(A2)'$, and $(A3)$ are satisfied. Then there exists a unique solution pair $(\lambda, r) \in L^2_{\mathcal{F}}(0, 1; H^1) \times [L^2_{\mathcal{F}}(0, 1; H^0)]^{d'}$, where $r := (r^1, r^2, \dots, r^{d'})$, of the following backward SPDE:

$$d\lambda(t) = - \left[A^*(t, U(t))\lambda(t) + \sum_{k=1}^{d'} M^{k*}(t, U(t))r^k(t) + F(t, U(t)) \right] dt$$

$$+ \sum_{k=1}^{d'} r^k(t) dW_k(t), \quad t \in [0, 1],$$

(4.1)

$$\lambda(1) = G.$$

Moreover, there is a constant N_7 that depend only on K and δ such that

$$E \int_0^1 \left[\|\lambda(t)\|_1^2 + \sum_{k=1}^{d'} \|r^k(t)\|_0^2 \right] dt \leq N_7 E \left[\int_0^1 \|F(t, U(t))\|_{-1}^2 dt + \|G\|_0^2 \right].$$

(4.2)

Remark 4.1. Solutions of (4.1) are defined in a way similar to those of the SPDE (2.6). More precisely, $(\lambda, r) \in L^2_{\mathcal{F}}(0, 1; H^1) \times [L^2_{\mathcal{F}}(0, 1; H^0)]^{d'}$ is called a solution pair of (4.1) if, for each $\eta \in C_0^\infty(R^d)$ and almost all $(t, \omega) \in [0, 1] \times \Omega$,

$$(\lambda(t), \eta)_0 = (G, \eta)_0 + \int_t^1 \left[(\lambda(s), A(s, U(s))\eta)_0 + \sum_{k=1}^{d'} (r^k(s), M^k(s, U(s))\eta)_0 \right. \\ \left. + (F(s, U(s)), \eta)_0 \right] ds - \sum_{k=1}^{d'} \int_t^1 (r^k(s), \eta)_0 dW_k(s).$$

So (4.1) can be regarded as in the H^{-1} space.

Proof of Theorem 4.1. To avoid notational complexity, we prove the theorem for $d' = 1$ (there is no essential difficulty when $d' > 1$). Thus the index “ k ” will be dropped. On the other hand, we also omit to write the control $U(t)$ since it is fixed.

Uniqueness. Suppose that $(\lambda, r) \in L^2_{\mathcal{F}}(0, 1; H^1) \times L^2_{\mathcal{F}}(0, 1; H^0)$ satisfies

$$d\lambda(t) = -[A^*(t)\lambda(t) + M^*(t)r(t)]dt + r(t)dW(t), \quad t \in [0, 1],$$

(4.3)

$$\lambda(1) = 0.$$

Consider the following auxiliary SPDE, which admits a unique solution $\rho \in L^2_{\mathcal{F}}(0, 1; H^1) \cap L^2(\Omega; C(0, 1; H^0))$ by virtue of Proposition 3.3:

$$\begin{aligned} d\rho(t) &= [A(t)\rho(t) + \lambda(t)]dt + [M(t)\rho(t) + r(t)]dW(t), \quad t \in [0, 1], \\ \rho(0) &= 0. \end{aligned}$$

Applying Ito's formula, we obtain

$$\begin{aligned} d(\rho(t), \lambda(t))_0 &= [\|\lambda(t)\|_0^2 + \|r(t)\|_0^2]dt \\ &\quad + [(M(t)\rho(t) + r(t), \lambda(t))_0 + (\rho(t), r(t))_0]dW(t). \end{aligned}$$

Hence, $E \int_0^1 [\|\lambda(t)\|_0^2 + \|r(t)\|_0^2]dt = 0$. This implies the uniqueness.

Before proceeding to the proof of the existence, we introduce an adjoint equation in finite dimension, which was originally obtained by Bismut [5]; see also [2], [3].

LEMMA 4.1. *Let n be a fixed positive integer. Suppose that we are given $A_n, M_n \in L^\infty_{\mathcal{F}}(0, 1; R^{n \times n})$, $F_n \in L^2_{\mathcal{F}}(0, 1; R^n)$, and $G_n \in R^n$. Then there exists uniquely a pair $(\lambda_n, r_n) \in L^2_{\mathcal{F}}(0, 1; R^n) \times L^2_{\mathcal{F}}(0, 1; R^n)$ satisfying the following backward SDE:*

$$(4.4) \quad \begin{aligned} d\lambda_n(t) &= -[A_n^T(t)\lambda_n(t) + M_n^T(t)r_n(t) + F_n(t)]dt + r_n(t)dW(t), \quad t \in [0, 1], \\ \lambda_n(1) &= G_n. \end{aligned}$$

Let us now continue proving Theorem 4.1.

Existence. Consider the Gelfand triple $H^1 \hookrightarrow H^0 \hookrightarrow H^{-1}$. Let $e_1, e_2, \dots, e_n, \dots$, be a Hilbert basis of H^1 , which is orthonormal as a basis of H^0 .

Fix a positive integer n . By Lemma 4.1, there is a unique pair

$$\tilde{\lambda}_n := (\lambda_{n1}, \lambda_{n2}, \dots, \lambda_{nn})^T \in L^2_{\mathcal{F}}(0, 1; R^n)$$

and

$$\tilde{r}_n := (r_{n1}, r_{n2}, \dots, r_{nn})^T \in L^2_{\mathcal{F}}(0, 1; R^n)$$

satisfying

$$(4.5) \quad \begin{aligned} d\lambda_{ni}(t) &= - \left[\sum_{j=1}^n \langle e_j, A(t)e_i \rangle_0 \lambda_{nj}(t) + \sum_{j=1}^n \langle e_j, M(t)e_i \rangle_0 r_{nj}(t) \right. \\ &\quad \left. + \langle F(t), e_i \rangle_0 \right] dt + r_{ni}(t)dW(t), \quad t \in [0, 1], \\ \lambda_{ni}(1) &= G_{ni}, \quad i = 1, 2, \dots, n, \end{aligned}$$

where $\sum_{i=1}^n G_{ni}e_i := G_n \rightarrow G$ in H^0 as $n \rightarrow \infty$. Define

$$(4.6) \quad \lambda_n := \sum_{i=1}^n \lambda_{ni}e_i \in L^2_{\mathcal{F}}(0, 1; H^1)$$

and

$$(4.7) \quad r_n := \sum_{i=1}^n r_{ni}e_i \in L^2_{\mathcal{F}}(0, 1; H^1).$$

Applying Ito's formula to (4.5) and adding up in i from 1 to n , we have

$$(4.8) \quad d\|\lambda_n(t)\|_0^2 = -2[\langle \lambda_n(t), A(t)\lambda_n(t) \rangle_0 + (r_n(t), M(t)\lambda_n(t))_0 + \langle F(t), \lambda_n(t) \rangle_0]dt \\ + 2(r_n(t), \lambda_n(t))_0 dW(t) + \|r_n(t)\|_0^2 dt.$$

Hence,

$$(4.9) \quad E\|\lambda_n(t)\|_0^2 \\ = E\|G_n\|_0^2 + 2E \int_t^1 [\langle A(s)\lambda_n(s), \lambda_n(s) \rangle_0 + (r_n(s), M(s)\lambda_n(s))_0 \\ + \langle F(s), \lambda_n(s) \rangle_0 - \frac{1}{2} \cdot \|r_n(s)\|_0^2] ds \\ \leq E\|G_n\|_0^2 + E \int_t^1 [2\langle A(s)\lambda_n(s), \lambda_n(s) \rangle_0 + \|M(s)\lambda_n(s)\|_0^2 + 2\langle F(s), \lambda_n(s) \rangle_0] ds \\ \leq E\|G_n\|_0^2 + E \int_t^1 [-\delta\|\lambda_n(s)\|_1^2 + N_3\|\lambda_n(s)\|_0^2 + 2/\delta \cdot \|F(s)\|_{-1}^2 + \delta/2 \cdot \|\lambda_n(s)\|_1^2] ds \\ \leq -\delta/2 \cdot E \int_t^1 \|\lambda_n(s)\|_1^2 ds + E\|G_n\|_0^2 + N_8 E \int_t^1 [\|\lambda_n(s)\|_0^2 + \|F(s)\|_{-1}^2] ds.$$

So Gronwall's inequality yields

$$(4.10) \quad \sup_{0 \leq t \leq 1} E\|\lambda_n(t)\|_0^2 + \delta/2 \cdot E \int_0^1 \|\lambda_n(t)\|_1^2 dt \leq N_9 E \left(\int_0^1 \|F(t)\|_{-1}^2 dt + \|G_n\|_0^2 \right),$$

where N_9 depends only on K and δ .

Now let $\rho_n := (\rho_{n1}, \rho_{n2}, \dots, \rho_{nn})^T \in L^2_{\mathcal{F}}(0, 1; R^n)$ be the solution of the following SDE in R^n :

$$(4.11) \quad d\rho_{ni}(t) = \sum_{j=1}^n \langle A(t)e_j, e_i \rangle_0 \rho_{nj}(t) dt + \left[\sum_{j=1}^n (M(t)e_j, e_i)_0 \rho_{nj}(t) + r_{ni}(t) \right] dW(t), \\ \rho_{ni}(0) = 0, \quad i = 1, 2, \dots, n.$$

Define $\rho_n := \sum_{i=1}^n \rho_{ni} e_i \in L^2_{\mathcal{F}}(0, 1; H^1)$. By a calculation similar to the above, we have

$$(4.12) \quad E\|\rho_n(t)\|_0^2 = E \int_0^t [2\langle A(s)\rho_n(s), \rho_n(s) \rangle_0 + \|M(s)\rho_n(s) + r_n(s)\|_0^2] ds \\ \leq -\delta E \int_0^t \|\rho_n(s)\|_1^2 ds + N_3 E \int_0^t [\|\rho_n(s)\|_0^2 + \|r_n(s)\|_0^2] ds.$$

Applying again Gronwall's inequality, we obtain

$$(4.13) \quad \sup_{0 \leq t \leq 1} E\|\rho_n(t)\|_0^2 + \delta E \int_0^1 \|\rho_n(t)\|_1^2 dt \leq N_3 \exp(N_3) E \int_0^1 \|r_n(t)\|_0^2 dt.$$

On the other hand, Ito's formula gives

$$(4.14) \quad d \sum_{i=1}^n \lambda_{ni}(t) \rho_{ni}(t) = \sum_{i=1}^n \{ -[\langle A^*(t)\lambda_n(t), e_i \rangle_0 + (M^*(t)r_n(t), e_i)_0 \\ + \langle F(t), e_i \rangle_0] \rho_{ni}(t) + \lambda_{ni}(t) \langle A(t)\rho_n(t), e_i \rangle_0 \\ + r_{ni}(t) [(M(t)\rho_n(t), e_i)_0 + r_{ni}(t)] \} dt + \{ \dots \} dW(t) \\ = [\|r_n(t)\|_0^2 - \langle F(t), \rho_n(t) \rangle_0] dt + \{ \dots \} dW(t).$$

Integrating from 0 to 1 and taking expectation, we have

$$\begin{aligned}
 E \int_0^1 \|r_n(t)\|_0^2 dt &= E \left[\int_0^1 \langle F(t), \rho_n(t) \rangle_0 dt + (G_n, \rho_n(1))_0 \right] \\
 &\leq \left(E \int_0^1 \|F(t)\|_{-1}^2 dt \right)^{1/2} \left(E \int_0^1 \|\rho_n(t)\|_1^2 dt \right)^{1/2} \\
 &\quad + (E \|G_n\|_0^2)^{1/2} (E \|\rho_n(1)\|_0^2)^{1/2} \\
 (4.15) \quad &\leq \left(N_{10} E \int_0^1 \|r_n(t)\|_0^2 dt \right)^{1/2} \\
 &\quad \cdot \left\{ \left(E \int_0^1 \|F(t)\|_{-1}^2 dt \right)^{1/2} + (E \|G_n\|_0^2)^{1/2} \right\},
 \end{aligned}$$

where $N_{10} = \max\{N_3 \exp(N_3), 1/\delta \cdot N_3 \exp(N_3)\}$. Consequently,

$$(4.16) \quad E \int_0^1 \|r_n(t)\|_0^2 dt \leq 2N_{10} E \left[\int_0^1 \|F(t)\|_{-1}^2 dt + \|G_n\|_0^2 \right].$$

By (4.10) and (4.16), there exist a subsequence $\{n'\}$ of $\{n\}$ and a pair $(\lambda, r) \in L^2_{\mathcal{F}}(0, 1; H^1) \times L^2_{\mathcal{F}}(0, 1; H^0)$ such that

$$(4.17) \quad \lambda_{n'} \rightarrow \lambda \text{ weakly in } L^2([0, 1] \times \Omega; H^1)$$

and

$$(4.18) \quad r_{n'} \rightarrow r \text{ weakly in } L^2([0, 1] \times \Omega; H^0) \text{ as } n' \rightarrow \infty.$$

We show that (λ, r) satisfies (4.1). To this end, let γ be an absolutely continuous function from $[0, 1]$ to R^1 with $\dot{\gamma} := d\gamma/dt \in L^2[0, 1]$ and $\gamma(0) = 0$. Set $\gamma_i(t) := \gamma(t)e_i$. Multiplying (4.5) by $\gamma_i(t)$ and using Ito's formula, we have

$$\begin{aligned}
 &\int_0^t (\lambda_n(t), \dot{\gamma}_i(t))_0 dt + \int_0^1 (r_n(t), \gamma_i(t))_0 dW(t) \\
 &= (G_n, \gamma_i(1))_0 + \int_0^1 [\langle \lambda_n(t), A(t)\gamma_i(t) \rangle_0 + (r_n(t), M(t)\gamma_i(t))_0 + \langle F(t), \gamma_i(t) \rangle_0] dt.
 \end{aligned}$$

(4.19)

Letting n' go to infinity and observing (4.17) and (4.18), we conclude that

$$\begin{aligned}
 &\int_0^t (\lambda(t), \phi)_0 \dot{\gamma}(t) dt + \int_0^1 (r(t), \phi)_0 \gamma(t) dW(t) \\
 (4.20) \quad &= (G, \phi)_0 \gamma(1) + \int_0^1 [\langle \lambda(t), A(t)\phi \rangle_0 + (r(t), M(t)\phi)_0 + \langle F(t), \phi \rangle_0] \gamma(t) dt
 \end{aligned}$$

for any $\phi \in H^1$.

For any $t \in (0, 1)$, let

$$\gamma_\varepsilon(s) := \begin{cases} 0 & \text{if } s \leq t - \varepsilon/2, \\ 1/\varepsilon \cdot (s - t + \varepsilon/2) & \text{if } t - \varepsilon/2 < s < t + \varepsilon/2, \\ 1 & \text{if } s \geq t + \varepsilon/2. \end{cases}$$

Substituting (4.20) with γ_ε and letting $\varepsilon \rightarrow 0$, we arrive at

$$\begin{aligned} (\lambda(t), \phi)_0 + \int_t^1 (r(s), \phi)_0 dW(s) &= (G, \phi)_0 \\ &+ \int_t^1 [\langle \lambda(s), A(s)\phi \rangle_0 + (r(s), M(s)\phi)_0 \\ &+ \langle F(s), \phi \rangle_0] ds \\ &\text{for any } \phi \in H^1, \text{ a.e. } t \in [0, 1]. \end{aligned}$$

This implies that (λ, r) satisfies (4.1). Finally, (4.2) is obtained by letting $n' \rightarrow \infty$ in (4.10) and (4.16). \square

COROLLARY 4.1. *Let the same assumptions as in Theorem 4.1 be satisfied. Given $\tilde{f} \in L^2_{\mathcal{F}}(0, 1; H^{-1})$ and $\tilde{g}^k \in L^2_{\mathcal{F}}(0, 1; H^0)$, $k = 1, 2, \dots, d'$, suppose that*

$$\xi \in L^2_{\mathcal{F}}(0, 1; H^1) \cap L^2(\Omega; C(0, 1; H^0))$$

satisfies

$$d\xi(t) = [A(t, U(t))\xi(t) + \tilde{f}(t)]dt + [M^k(t, U(t))\xi(t) + \tilde{g}^k(t)]dW_k(t), \quad t \in [0, 1],$$

(4.21)

and that (λ, r) satisfies (4.1). Then, for any $[\alpha, \beta] \subset [0, 1]$,

$$\begin{aligned} (4.22) \quad &E \left[\int_\alpha^\beta \langle F(t, U(t)), \xi(t) \rangle_0 dt + (\lambda(\beta), \xi(\beta))_0 \right] \\ &= E \left\{ \int_\alpha^\beta \left[\langle \lambda(t), f(t) \rangle_0 + \sum_{k=1}^{d'} (r^k(t), g^k(t))_0 \right] dt + (\lambda(\alpha), \xi(\alpha))_0 \right\}. \end{aligned}$$

Proof. The result is easily derived by applying Ito's formula to $(\lambda(t), \xi(t))_0$. \square

Remark 4.2. In Theorem 4.1 and Corollary 4.1, condition (A1)₀ can be weakened by assuming that all the coefficients a^{ij} , and so forth, are only bounded measurable in x . However, under such a weaker condition, the adjoint operators A^* and M^{k*} can be no longer written explicitly as (2.3) and (2.4), respectively. Instead, as operators mapping H^1 into H^{-1} , they can be determined by the following formulae:

$$\langle A^*(t, u)\phi, \psi \rangle_0 := \langle \phi, A(t, u)\psi \rangle_0 \text{ and}$$

$$\langle M^{k*}(t, u)\phi, \psi \rangle_0 := \langle \phi, M^k(t, u)\psi \rangle_0 \text{ for } \phi, \psi \in H^1.$$

It should be also noted that the definition of solutions of (4.1) does not require explicit expressions of A^* and M^{k*} ; see Remark 4.1.

5. Necessary conditions of optimality. We study in this section necessary conditions of an optimal control for the general system (2.6) with the cost functional (2.8).

THEOREM 5.1. *Assume that (A1)₂, (A2)', (A3), (A4), and (A5) are satisfied and that $(\Omega, \mathcal{F}, P, W, \hat{U})$ is an optimal control along with the corresponding optimal state \hat{q} . Then, for almost every $t \in [0, 1]$, we have the maximum condition*

$$(5.1) \quad H(t, \hat{q}(t), \hat{U}(t), \lambda(t), r(t)) = \max_{u \in I'} H(t, \hat{q}(t), u, \lambda(t), r(t)), \quad P - \text{a.s.},$$

where (λ, r) is the solution pair of (4.1) with $U(t) \equiv \hat{U}(t)$, and the Hamiltonian H is defined by

$$\begin{aligned} H(t, \phi, u, \zeta, \eta) &:= -\langle A(t, u)\phi, \zeta \rangle_0 - \langle f(t, u), \zeta \rangle_0 \\ &\quad - \sum_{k=1}^{d'} [(M^k(t, u)\phi, \eta^k)_0 + (g^k(t, u), \eta^k)_0] - \langle F(t, u), \phi \rangle_0, \\ &\quad \text{for } (t, \phi, u, \zeta, \eta) \in [0, 1] \times H^1 \times \Gamma \times H^1 \times (H^0)^{d'}. \end{aligned} \quad (5.2)$$

Proof. We assume that $d' = 1$ and set $\hat{A}(t) := A(t, \hat{U}(t))$, and so on, for simplicity. From Proposition 3.3, it follows that $\hat{q} \in L^2_{\mathcal{F}}(0, 1; H^2) \cap L^2(\Omega; C(0, 1; H^1))$. Hence $\hat{q}(t) \in L^2(\Omega; H^2)$ for almost every $t \in [0, 1]$. Fix a time $\bar{t} \in [0, 1]$ such that $\hat{q}(\bar{t}) \in L^2(\Omega; H^2)$, along with a Γ -valued, $\mathcal{F}_{\bar{t}}$ -measurable random variable u . For any $\varepsilon \in (0, 1 - \bar{t})$, define $U_\varepsilon \in U_{ad}$ by

$$U_\varepsilon(t) := \begin{cases} u, & t \in [\bar{t}, \bar{t} + \varepsilon], \\ \hat{U}(t), & t \in [0, 1] \setminus [\bar{t}, \bar{t} + \varepsilon]. \end{cases}$$

Let q_ε be the response for U_ε , namely,

$$q_\varepsilon(t) = \hat{q}(t), \quad t \in [0, \bar{t}], \quad (5.3)$$

$$q_\varepsilon(t) = \hat{q}(\bar{t}) + \int_{\bar{t}}^t [A(s, u)q_\varepsilon(s) + f(s, u)]ds + \int_{\bar{t}}^t [M(s, u)q_\varepsilon(s) + g(s, u)]dW(s), \quad t \in [\bar{t}, \bar{t} + \varepsilon], \quad (5.4)$$

and

$$q_\varepsilon(t) = q_\varepsilon(\bar{t} + \varepsilon) + \int_{\bar{t} + \varepsilon}^t [\hat{A}(s)q_\varepsilon(s) + \hat{f}(s)]ds + \int_{\bar{t} + \varepsilon}^t [\hat{M}(s)q_\varepsilon(s) + \hat{g}(s)]dW(s), \quad t \in [\bar{t} + \varepsilon, 1]. \quad (5.5)$$

It follows from Proposition 3.3 that $q_\varepsilon \in L^2(\Omega; C(\bar{t}, 1; H^2))$ and

$$\begin{aligned} E \left(\sup_{\bar{t} \leq t \leq \bar{t} + \varepsilon} \|q_\varepsilon(t)\|_2^2 \right) &\leq N_6 E \left\{ \|\hat{q}(\bar{t})\|_2^2 + \int_{\bar{t}}^{\bar{t} + \varepsilon} [\|f(t, u)\|_1^2 + \|g(t, u)\|_2^2] dt \right\} \\ &\leq N_6 E(\|\hat{q}(\bar{t})\|_2^2 + 2K^2\varepsilon). \end{aligned} \quad (5.6)$$

Define $\xi_\varepsilon(t) := q_\varepsilon(t) - \hat{q}(t)$ for $t \in [0, 1]$. Then ξ_ε satisfies

$$\begin{aligned} d\xi_\varepsilon(t) &= [\hat{A}(t)\xi_\varepsilon(t) + (A(t, u) - \hat{A}(t))q_\varepsilon(t) + f(t, u) - \hat{f}(t)]dt \\ &\quad + [\hat{M}(t)\xi_\varepsilon(t) + (M(t, u) - \hat{M}(t))q_\varepsilon(t) + g(t, u) - \hat{g}(t)]dW(t), \\ &\quad t \in [\bar{t}, \bar{t} + \varepsilon] \end{aligned} \quad (5.7)$$

and

$$d\xi_\varepsilon(t) = \hat{A}(t)\xi_\varepsilon(t)dt + \hat{M}(t)\xi_\varepsilon(t)dW(t), \quad t \in [\bar{t} + \varepsilon, 1]. \quad (5.8)$$

Since \hat{U} is optimal, we have

$$\begin{aligned}
 0 \leq J(U_\varepsilon) - J(\hat{U}) &= E \left\{ \int_{\bar{t}}^1 [\langle F(t, U_\varepsilon(t)), q_\varepsilon(t) \rangle_0 - \langle \hat{F}(t), \hat{q}(t) \rangle_0] dt \right. \\
 &\quad \left. + (G, q_\varepsilon(1) - \hat{q}(1))_0 \right\} \\
 (5.9) \quad &= E \int_{\bar{t}}^{\bar{t}+\varepsilon} [\langle F(t, u) - \hat{F}(t), q_\varepsilon(t) \rangle_0 + \langle \hat{F}(t), \xi_\varepsilon(t) \rangle_0] dt \\
 &\quad + E \left[\int_{\bar{t}+\varepsilon}^1 \langle \hat{F}(t), \xi_\varepsilon(t) \rangle_0 dt + (G, \xi_\varepsilon(1))_0 \right] = I_1 + I_2.
 \end{aligned}$$

Applying Corollary 4.1 to (5.8) and (4.1), we have $I_2 = E(\lambda(\bar{t} + \varepsilon), \xi_\varepsilon(\bar{t} + \varepsilon))_0$. Thus we can rewrite (5.9) as

$$\begin{aligned}
 0 \leq E \int_{\bar{t}}^{\bar{t}+\varepsilon} \langle F(t, u) - \hat{F}(t), q_\varepsilon(t) \rangle_0 dt \\
 (5.10) \quad + E \left[\int_{\bar{t}}^{\bar{t}+\varepsilon} \langle \hat{F}(t), \xi_\varepsilon(t) \rangle_0 dt + (\lambda(\bar{t} + \varepsilon), \xi_\varepsilon(\bar{t} + \varepsilon))_0 \right].
 \end{aligned}$$

Applying Corollary 4.1 again to (5.7) and (4.1), we obtain

$$\begin{aligned}
 0 \leq E \int_{\bar{t}}^{\bar{t}+\varepsilon} [\langle F(t, u) - \hat{F}(t), q_\varepsilon(t) \rangle_0 + \langle \lambda(t), (A(t, u) - \hat{A}(t))q_\varepsilon(t) \rangle_0 \\
 (5.11) \quad + (\lambda(t), f(t, u) - \hat{f}(t))_0 \\
 + (r(t), (M(t, u) - \hat{M}(t))q_\varepsilon(t) + g(t, u) - \hat{g}(t))_0] dt.
 \end{aligned}$$

On the other hand, we have

$$\begin{aligned}
 (1/\varepsilon) \int_{\bar{t}}^{\bar{t}+\varepsilon} E \langle \lambda(t), (A(t, u) - \hat{A}(t))(q_\varepsilon(t) - \hat{q}(t)) \rangle_0 dt \\
 (5.12) \quad \leq \text{const } (1/\varepsilon) \int_{\bar{t}}^{\bar{t}+\varepsilon} E(\|\lambda(t)\|_1 \|\xi_\varepsilon(t)\|_1) dt.
 \end{aligned}$$

Recall that ξ_ε satisfies (5.7) on $[\bar{t}, \bar{t} + \varepsilon]$ with $\xi_\varepsilon(\bar{t}) = 0$. Hence Proposition 3.3 gives

$$\begin{aligned}
 E \left(\sup_{\bar{t} \leq t \leq \bar{t}+\varepsilon} \|\xi_\varepsilon(t)\|_1^2 \right) \\
 (5.13) \quad \leq N_6 E \int_{\bar{t}}^{\bar{t}+\varepsilon} [\|(A(t, u) - \hat{A}(t))q_\varepsilon(t) + f(t, u) - \hat{f}(t)\|_0^2 \\
 + \|(M(t, u) - \hat{M}(t))q_\varepsilon(t) + g(t, u) - \hat{g}(t)\|_1^2] dt \\
 \leq \text{const } E \int_{\bar{t}}^{\bar{t}+\varepsilon} (\|q_\varepsilon(t)\|_2^2 + 1) dt \\
 \leq \text{const } (E\|\hat{q}(\bar{t})\|_2^2 + 1)\varepsilon \quad (\text{by (5.6)}) \\
 \leq \text{const } \varepsilon \quad (\text{Note that we have fixed } \bar{t} \text{ such that } E\|\hat{q}(\bar{t})\|_2^2 < +\infty).
 \end{aligned}$$

Thus, (5.12) reduces to

$$\begin{aligned}
 (1/\varepsilon) \int_{\bar{t}}^{\bar{t}+\varepsilon} E \langle \lambda(t), (A(t, u) - \hat{A}(t))(q_\varepsilon(t) - \hat{q}(t)) \rangle_0 dt \\
 (5.14) \quad \leq \text{const } (1/\varepsilon) \int_{\bar{t}}^{\bar{t}+\varepsilon} \{(\varepsilon^{1/3}/2)E\|\lambda(t)\|_1^2 + [1/(2\varepsilon^{1/3})]E\|\xi_\varepsilon(t)\|_1^2\} dt \\
 \leq \text{const } \left[\varepsilon^{1/3} \cdot (1/\varepsilon) \int_{\bar{t}}^{\bar{t}+\varepsilon} E\|\lambda(t)\|_1^2 dt + (1/\varepsilon)\varepsilon(1/\varepsilon^{1/3})\varepsilon \right] \\
 \rightarrow 0 \text{ as } \varepsilon \rightarrow 0,
 \end{aligned}$$

provided that \bar{t} is a Lebesgue point of the function $t \rightarrow E\|\lambda(t)\|_1^2$. Similarly, we have

$$(5.15) \quad \begin{aligned} (1/\varepsilon)E \int_{\bar{t}}^{\bar{t}+\varepsilon} [\langle F(t, u) - \hat{F}(t), q_\varepsilon(t) - \hat{q}(t) \rangle_0 \\ + \langle r(t), (M(t, u) - \hat{M}(t))(q_\varepsilon(t) - \hat{q}(t)) \rangle_0] dt \\ \rightarrow 0 \text{ as } \varepsilon \rightarrow 0, \end{aligned}$$

provided that \bar{t} is a Lebesgue point of the function $t \rightarrow E\|r(t)\|_0^2$. Thus (5.11) becomes

$$(5.16) \quad \begin{aligned} 0 \leq E \int_{\bar{t}}^{\bar{t}+\varepsilon} [\langle F(t, u) - \hat{F}(t), \hat{q}(t) \rangle_0 + \langle \lambda(t), (A(t, u) - \hat{A}(t))\hat{q}(t) \rangle_0 \\ + \langle \lambda(t), f(t, u) - \hat{f}(t) \rangle_0 + \langle r(t), (M(t, u) - \hat{M}(t))\hat{q}(t) \\ + g(t, u) - \hat{g}(t) \rangle_0] dt + o(\varepsilon). \end{aligned}$$

Dividing (5.16) by ε and letting $\varepsilon \rightarrow 0$, we obtain

$$(5.17) \quad EH(\bar{t}, \hat{q}(\bar{t}), \hat{U}(\bar{t}), \lambda(\bar{t}), r(\bar{t})) \geq EH(\bar{t}, \hat{q}(\bar{t}), u, \lambda(\bar{t}), r(\bar{t})).$$

Therefore, the desired result (5.1) follows from a standard argument; see, for example, [11]. This concludes the theorem. \square

Remark 5.1. In Theorem 5.1, all the coefficients appearing in the control problem are allowed to depend on the control variable.

Remark 5.2. By the above proof (especially the argument between (5.12) and (5.17)), it is easy to see that, if a^{ij} and σ^{ik} contain no control variable, then the assumptions of Theorem 5.1 can be considerably relaxed. More precisely, (A1)₂ can be replaced by the assumption that all the coefficients a^{ij} , and so forth, are bounded measurable in x , and (A4) and (A5) together can be replaced by the assumptions that $q_0 \in H^0$, $f(t, u) \in H^{-1}$, $g^k(t, u) \in H^0$, and that their respective norms are bounded. Therefore, all the regularity restrictions imposed in Bensoussan [4] can be removed. It should be also noted that M^k is allowed to be unbounded in our results, compared with [4].

6. Discussion and application.

6.1. Degenerate cases. The main results in §§4 and 5 are derived under assumption (A2)'; namely, system (2.6) is nondegenerate. There seems to be some essential difficulties in treating possibly degenerate systems, and the adjoint equation (4.1) may no longer admit a solution pair, no matter how high regularity assumptions to be imposed on the functions F and G . At least the method of proving Theorem 4.1 no longer applies. Indeed, the basic idea behind the proof of Theorem 4.1 is to approximate the infinite-dimensional equation by certain finite-dimensional equations and to use estimate (3.5) to prove that the approximate solutions are weakly compact. However, when the systems are possibly degenerate, we have only estimate (3.1) available, which could not ensure the above compactness. On the other hand, estimate (3.1) is the "best one" in that it cannot be further improved.

However, our approach still works in the degenerate case if M^k is of order zero, for which a stronger estimate (3.2) is available.

Let us now introduce the following assumptions:

(B1) $F(t, u) \in H^1$, $G \in H^1$, and $\|F(t, u)\|_1 + \|G\|_1 \leq K$;

(B2) $f, g^k : [0, 1] \times R^d \times \Gamma \rightarrow R^1$ are measurable in (t, x, u) and continuous in $u, k = 1, 2, \dots, d'$. Furthermore, $f(t, \cdot, u), g^k(t, \cdot, u) \in H^3$; and

$$|f(t, x, u)| + |g^k(t, x, u)| + \|f(t, \cdot, u)\|_3 + \|g^k(t, \cdot, u)\|_3 \leq K, k = 1, 2, \dots, d';$$

(B3) $q_0 \in H^3$.

THEOREM 6.1. Assume that (B1), (A1)₁, and (A2) are satisfied with $\sigma^{ik} = 0$. Then, for any $(\Omega, \mathcal{F}, P, W, U) \in U_{ad}$, there exists a unique solution pair $(\lambda, r) \in L^2_{\mathcal{F}}(0, 1; H^1) \times [L^2_{\mathcal{F}}(0, 1; H^1)]^{d'}$ of the adjoint equation (4.1). Moreover, there is a constant N_{11} that depends only on K such that

$$(6.1) \quad E \int_0^1 \left[\|\lambda(t)\|_1^2 + \sum_{k=1}^{d'} \|r^k(t)\|_1^2 \right] dt \leq N_{11} E \left[\int_0^1 \|F(t, U(t))\|_1^2 dt + \|G\|_1^2 \right].$$

Proof. Once estimate (3.2) is observed, the theorem can be proved in a way similar to that of Theorem 4.1. The proof is left to the reader. \square

THEOREM 6.2. Assume that (B1)–(B3), (A1)₃, and (A2) are satisfied with $\sigma^{ik} = 0$ and that $(\Omega, \mathcal{F}, P, W, \hat{U}) \in U_{ad}$ is an optimal control along with the optimal state \hat{q} . Then for almost every $t \in [0, 1]$,

$$(6.2) \quad H(t, \hat{q}(t), \hat{U}(t), \lambda(t), r(t)) = \max_{u \in \Gamma} H(t, \hat{q}(t), u, \lambda(t), r(t)), \quad P - \text{a.s.},$$

where (λ, r) is the solution pair of (4.1) with the control $U = \hat{U}$, and the Hamiltonian H is defined by (5.2).

Proof. Basing on Theorem 6.1, we only need to slightly modify the proof of Theorem 5.1. From Corollary 3.3, it follows that $\hat{q} \in L^2_{\mathcal{F}}(0, 1; H^3) \cap L^2(\Omega; C(0, 1; H^2))$. Hence $\hat{q}(t) \in L^2(\Omega; H^3)$ for almost every $t \in [0, 1]$. Fix $\bar{t} \in [0, 1)$ such that $\hat{q}(\bar{t}) \in L^2(\Omega; H^3)$. For $\varepsilon \in (0, 1 - \bar{t})$, define q_ε by (5.3)–(5.5). By virtue of Corollary 3.3, we have

$$\begin{aligned} E \left(\sup_{\bar{t} \leq t \leq \bar{t} + \varepsilon} \|q_\varepsilon(t)\|_3^2 \right) &\leq N_4 E \left\{ \|\hat{q}(\bar{t})\|_3^2 + \int_{\bar{t}}^{\bar{t} + \varepsilon} [\|f(s, u)\|_3^2 + \|g(s, u)\|_3^2] ds \right\} \\ &\leq N_4 E (\|\hat{q}(\bar{t})\|_3^2 + 2K^2 \varepsilon). \end{aligned}$$

Define $\xi_\varepsilon(t) := q_\varepsilon(t) - \hat{q}(t)$. Then ξ_ε satisfies (5.7) on $[\bar{t}, \bar{t} + \varepsilon]$. Hence,

$$\begin{aligned} E \left(\sup_{\bar{t} \leq t \leq \bar{t} + \varepsilon} \|\xi_\varepsilon(t)\|_1^2 \right) &\leq N_4 E \int_{\bar{t}}^{\bar{t} + \varepsilon} [\|(A(t, u) - \hat{A}(t))q_\varepsilon(t) + f(t, u) - \hat{f}(t)\|_1^2 \\ &\quad + \|(M(t, u) - \hat{M}(t))q_\varepsilon(t) + g(t, u) - \hat{g}(t)\|_1^2] dt \\ &\leq \text{const } E \int_{\bar{t}}^{\bar{t} + \varepsilon} (\|q_\varepsilon(t)\|_3^2 + 1) dt \\ &\leq \text{const } (E \|\hat{q}(\bar{t})\|_3^2 + 1) \varepsilon \leq \text{const } \varepsilon. \end{aligned}$$

In what follows, we must only repeat the arguments in the proof of Theorem 5.1 to obtain the desired result. \square

Remark 6.1. Theorem 6.2 requires the higher regularity on the coefficients a^{ij} , f , and so forth, as well as on the initial state q_0 . However, if the second-order coefficients of the operator $A(t, u)$ contain no control variable, then it is easy to verify by the proof of Theorem 6.2 that all the regularity conditions can be reduced by two orders. It should be also noted that Theorem 6.2 extends the results of Bensoussan [4] to possibly degenerate systems.

6.2. Application to partially observed diffusions. The optimal control problem of partially observed diffusions with general nonlinear cost functionals can be formulated as a control problem of linear SPDEs (Zakai's equations) with linear cost functionals. Hence the results obtained in the previous sections can be applied directly to partially observed diffusions.

First, let us remark that the results in §§5 and 6 can be easily extended to the following class of systems:

$$(6.3) \quad \begin{aligned} dq(t) &= [A(t, W(t), U(t))q(t) + f(t, W(t), U(t))]dt \\ &\quad + [M^k(t, W(t), U(t))q(t) + g^k(t, W(t), U(t))]dW_k(t), \quad t \in [0, 1], \\ q(0) &= q_0, \end{aligned}$$

where

$$\begin{aligned} A(t, w, u)\phi(x) &:= \partial_i(a^{ij}(t, x, w, u)\partial_j\phi(x)) + b^i(t, x, w, u)\partial_i\phi(x) + c(t, x, w, u)\phi(x), \\ M^k(t, w, u)\phi(x) &:= \sigma^{ik}(t, x, w, u)\partial_i\phi(x) + h^k(t, x, w, u)\phi(x), \quad k = 1, 2, \dots, d'. \end{aligned}$$

Note only the continuity of the coefficients a^{ij} and so forth in $w \in R^{d'}$ will be needed later.

Let W and \tilde{W} be two independent Brownian motions on a probability space (Ω, \mathcal{F}, P) , with values in R^d and $R^{d'}$, respectively. Consider the following SDE in R^d :

$$(6.4) \quad \begin{aligned} dX(t) &= \gamma(t, X(t), Y(t), U(t))dt + \alpha(t, X(t), Y(t), U(t))dW(t) \\ &\quad + \sigma(t, X(t), Y(t), U(t))d\tilde{W}(t), \quad t \in [0, 1], \\ X(0) &= \xi, \end{aligned}$$

with the observation

$$(6.5) \quad \begin{aligned} dY(t) &= \kappa(t, X(t), Y(t), U(t))dt + d\tilde{W}(t), \quad t \in [0, 1], \\ Y(0) &= 0, \end{aligned}$$

where U is an admissible control, namely, $\{U(t) : 0 \leq t \leq 1\}$ is a Γ -valued, $\sigma\{Y(s) : 0 \leq s \leq t\}$ -adapted measurable process. Note σ is the correlation between the state and the observation noises.

Let $F : [0, 1] \times R^d \times R^{d'} \times \Gamma \rightarrow R^1$, $G : R^d \times R^{d'} \rightarrow R^1$ be given. The objective is to minimize the cost functional defined by

$$(6.6) \quad J(U) := E \left[\int_0^1 F(t, X(t), Y(t), U(t))dt + G(X(1), Y(1)) \right]$$

over the set of admissible controls.

Note that W and Y are independent Brownian motions under a new probability \tilde{P} defined by $d\tilde{P} := \rho^{-1}(1)dP$, where

$$\rho(t) := \exp \left[\int_0^t \kappa(s, X(s), Y(s), U(s))dY(s) - \frac{1}{2} \int_0^t |\kappa(s, X(s), Y(s), U(s))|^2 ds \right].$$

Consider the following SPDE:

$$(6.7) \quad \begin{aligned} dq(t) &= A(t, Y(t), U(t))q(t)dt + M^k(t, Y(t), U(t))q(t)dY_k(t), \quad t \in [0, 1], \\ q(0) &= q_0, \end{aligned}$$

where $Y_k, k = 1, 2, \dots, d'$, are the components of Y , and

$$\begin{aligned} A(t, y, u)\phi(x) &:= \partial_i(a^{ij}(t, x, y, u)\partial_j\phi(x)) - \partial_i(a^i(t, x, y, u)\phi(x)), \\ M^k(t, y, u)\phi(x) &:= -\sigma^{ik}(t, x, y, u)\partial_i\phi(x) + h^k(t, x, y, u)\phi(x), \\ (a^{ij}(t, x, y, u))_{ij} &\equiv a(t, x, y, u) := [\sigma\sigma^T(t, x, y, u) + \alpha\alpha^T(t, x, y, u)]/2, \\ a^i(t, x, y, u) &:= \gamma^i(t, x, y, u) - \partial_j a^{ij}(t, x, y, u), \\ h^k(t, x, y, u) &:= \kappa^k(t, x, y, u) - \partial_i \sigma^{ik}(t, x, y, u), \quad i, j = 1, 2, \dots, d \quad k = 1, 2, \dots, d', \\ q_0 &:= \text{the density of } \xi. \end{aligned}$$

Then $q(t)$ proves to be the unnormalized conditional probability density of the state process $X(t)$, and the cost functional (6.6) reduces to

$$(6.8) \quad J(U) = \tilde{E} \left[\int_0^1 (F(t, \cdot, Y(t), U(t)), q(t))_0 dt + (G(\cdot Y(1)), q(1))_0 \right];$$

see Pardoux [13], Rozovskii [14], and Nagase and Nisio [12] for details.

Now we have turned problem (6.4)–(6.6) to the one already solved in §§5 and 6, observing the remark at the beginning of this section. So we can obtain the necessary conditions for an optimal solution to problem (6.4)–(6.6) by appropriately interpreting the assumptions and conclusions. Here we omit this simple interpretative work and only remark that, in the present case, (A2) is satisfied automatically; (A2)' is also satisfied, provided that $\alpha\alpha^T$ is uniformly positive definite.

7. Concluding remarks. In this paper, a finite-dimensional approximation approach and some a priori estimates of differential operators are employed to solve the adjoint equations of linear SPDEs, based on which necessary conditions of optimality for controlled SPDEs are derived. For some sufficient conditions for the existence of an optimal control, refer to [12], [15].

The adjoint equations of SPDEs may also be applied to some other important problems besides the necessity of optimality. For example, a relationship between the adjoint process λ and the value function is given in [16], which enables us to gain some insight into the deep connection between the adjoint equations and the dynamic programming equations (Hamilton-Jacobi-Bellman equations). Further research on the analytical and qualitative properties of solutions of the adjoint equations together with their applications to linear SPDEs is carried out in [17].

Let us conclude the paper by noting that our results on the adjoint equations are by no means optimal. In view of the observation at the beginning of §6.1, it remains a challenging open problem to solve the adjoint equations of degenerate SPDEs in which diffusion terms contain first-order differential operators.

Acknowledgments. This research was carried out when I was visiting Kobe University. I would like to express my hearty thanks to Professor M. Nisio for her constant encouragement and kind hospitality. Thanks are also due to the referees for their helpful comments and suggestions.

REFERENCES

- [1] J. S. BARAS, R. J. ELLIOTT, AND M. KOHLMANN, *The partially observed stochastic minimum principle*, SIAM J. Control Optim., 27 (1989), pp. 1279–1292.
- [2] A. BENSOUSSAN, *Lecture on stochastic control, Part I*, Lecture Notes in Math., 972 (1983), pp. 1–39.

- [3] ———, *Stochastic maximum principle for distributed parameter system*, J. Franklin Inst., 315 (1983), pp. 387–406.
- [4] ———, *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169–222.
- [5] J. M. BISMUT, *Analyse convexe et probabilités*, Thèse, Faculté des Sciences de Paris, 1973.
- [6] G. DA PRATO, M. IANNELLI, AND L. TUBARO, *Some results on linear stochastic differential equations in Hilbert spaces*, Stochastics, 6 (1982), pp. 105–116.
- [7] U. G. HAUSSMANN, *The maximum principle for optimal control of diffusions with partial information*, SIAM J. Control Optim., 25 (1987), pp. 341–361.
- [8] N. V. KRYLOV AND B. L. ROZOVSKII, *On the Cauchy problem for linear stochastic partial differential equations*, Izv. Akad. Nauk SSSR Ser. Mat., 41 (1977), pp. 1329–1347; Mat. USSR-Izv., 11 (1977), pp. 1267–1284.
- [9] ———, *On characteristics of the degenerate parabolic Ito equations of the second order*, Proc. Petrovskii Sem., 8 (1982), pp. 153–168. (In Russian.)
- [10] ———, *Stochastic partial differential equations and diffusion processes*, Uspekhi Mat. Nauk, 37:6 (1982), pp. 75–95; Russian Math. Surveys, 37:6 (1982), pp. 81–105.
- [11] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control, 10 (1972), pp. 550–565.
- [12] N. NAGASE AND M. NISIO, *Optimal controls for stochastic partial differential equations*, SIAM J. Control Optim., 28 (1990), pp. 186–213.
- [13] E. PARDOUX, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–167.
- [14] B. L. ROZOVSKII, *Nonnegative L^1 -solutions of second order stochastic parabolic equations with random coefficients*, Steklov Seminar, Stat. Control of Stoch. Proc., 1984; Transl. Math. English, 1985, pp. 410–427.
- [15] X. Y. ZHOU, *On the existence of optimal relaxed controls for stochastic partial differential equations*, SIAM J. Control Optim., 30 (1992), pp. 247–261.
- [16] ———, *Remarks on optimal controls of stochastic partial differential equations*, System Control Lett., 16 (1991), pp. 465–472.
- [17] ———, *A duality analysis on stochastic partial differential equations*, J. Funct. Anal., 103 (1992), pp. 275–293.

TRACE REGULARITY IN THE BOUNDARY CONTROL OF A WAVE EQUATION*

JONG UHN KIM†

Abstract. The regularity of the boundary control of a wave equation when the control is obtained by the methods of Russell and Lagnese is investigated.

Key words. trace regularity, boundary control, wave equation, wave front

AMS subject classifications. 35L05, 35B37, 35B65, 93B05, 93C20

Introduction. In this paper we will discuss the regularity of the boundary control of a wave equation when the boundary control is obtained through the methods of Russell [16], [17], and Lagnese [8], [9].

For a space dimension that is an odd integer larger than one, Russell [16] applied Huyghen's principle directly to obtain the boundary control of a wave equation. For a space dimension that is even, he set forth a mathematical framework where the question of exact boundary controllability is reduced to the invertibility of a certain linear operator. He also showed that this can be achieved by the decay of local energy; however, the use of energy decay requires sufficiently large control time. This was improved by Lagnese [8] who used a different method to prove the invertibility, provided that the control time is larger than the diameter of the space domain. In fact, this is the optimal control time. Littman [14], [15] also made a contribution to this approach for exact boundary controllability of hyperbolic equations.

The above methods have been overshadowed by the more recent Hilbert uniqueness method. However, the above methods give rise to a very interesting analysis problem on the regularity of boundary control. Our purpose is to investigate this problem. This makes it necessary to study the trace regularity of solutions to a general second-order differential equation, which is interesting in its own light. This is done in §1 and the main result is Theorem 1.1. In §2, we apply this to the regularity of the boundary control when the control is obtained according to [8] and [16] under the assumption that the initial data have compact support. Our result in §1 does not apply to the boundary control obtained in [15], which we will not discuss in this paper.

As a simple byproduct, we also have application to the trace regularity of solutions to a hyperbolic equation with the Neumann boundary condition when the space domain is a half space and the initial data have compact support. If the coefficients of the equation are constants, this problem has been already discussed by Symes [18] and Lasiecka and Triggiani [10] using different methods. In §3, we extend their results to the case of variable coefficients which satisfy some restrictive conditions.

After the first version of this paper was completed, Professor I Lasiecka informed me of the work of Bao and Symes [2]. They established a result similar to our Theorem 1.1 in a more general setting. But our analysis is more elementary from a technical viewpoint and is applicable to the boundary control of a wave equation in a more straightforward manner.

Our method is microlocal analysis. In particular, Hörmander's result [5] on the propagation of singularities plays a crucial role. Microlocal analysis has become a powerful tool in control theory since the work of Bardos, Lebeau, and Rauch [4]. Using microlocal analysis, they obtained sharp conditions for the boundary control and stabilization of a wave equation. Its utility has been also manifested in [2], [3], [7], and [11].

* Received by the editors February 12, 1992; accepted for publication (in revised form) June 26, 1992.

† Department of Mathematics, Virginia Polytechnic Institute, Blacksburg, Virginia 24061.

Finally we remark that when the support of initial data is not compact, the question addressed in this paper is still open, which we hope to investigate in the future.

Notation. Let Ω be an open subset of R^N . $\mathcal{D}'(\Omega)$ is the space of distributions over Ω and $H^s(\Omega)$, and $s \in R$ denotes a Sobolev space as defined in [13]. $H_c^s(\Omega)$ is the space of elements of $H^s(\Omega)$ which have compact support in Ω . $H_{\text{loc}}^s(\Omega)$ denotes the space of distributions u in Ω such that $\psi u \in H^s(\Omega)$ for each $\psi \in C_0^\infty(\Omega)$. If Ω is a C^∞ manifold with boundary, $H^s(\Omega)$ and $H_{\text{loc}}^s(\Omega)$ are defined by means of coordinate patches and a partition of unity. For $u \in \mathcal{D}'(\Omega)$, we use the following notation: $\text{supp } u$ = the support of u , $\text{sing supp } u$ = the singular support of u , $WF(u)$ = the wave front set of u . D_{x_i} and D_y stand for $-i \partial / \partial x_i$ and $-i \partial / \partial y$, respectively, and $D_x^\alpha = D_{x_1}^{\alpha_1} \cdots D_{x_N}^{\alpha_N}$, for $\alpha = (\alpha_1, \dots, \alpha_N)$. If m is a real number, then

$$S_{1,0}^m(R^N) = \text{the set of all } a(x, \xi) \in C^\infty(R^N \times R^N),$$

which satisfies

$$|\partial_\xi^\alpha \partial_x^\beta a(x, \xi)| \leq C_{\alpha, \beta} (1 + |\xi|)^{m - |\alpha|},$$

for all $(x, \xi) \in R^N \times R^N$, for some positive constant $C_{\alpha, \beta}$. Here,

$$\partial_\xi^\alpha = \left(\frac{\partial}{\partial \xi_1} \right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial \xi_N} \right)^{\alpha_N} \quad \text{and} \quad |\alpha| = \alpha_1 + \cdots + \alpha_N.$$

Next we set

$$OPS^m(R^N) = \text{the set of all pseudodifferential operators of order } m \\ \text{whose symbols belong to } S_{1,0}^m(R^N).$$

If $p \in OPS^m(R^N)$, then

$$ES(p) = \text{the essential support of } p.$$

For its definition, see [19].

When $p \in OPS^{m_1}(R^N)$ and $q \in OPS^{m_2}(R^N)$, we write $[p, q] = pq - qp$. Finally, if the two sets G_1 and G_2 are disjoint, we write $G_1 \cap G_2 = \phi$.

1. Trace regularity of solutions to a second-order differential operator. In this section, Ω is an open subset of R^{n+1} and (y, x) denotes the variables in R^{n+1} with $y \in R$ and $x \in R^n$. We denote the dual variables by (η, ξ) with $\eta \in R$ and $\xi \in R^n$. We will consider a second-order differential operator of the form

$$(1.1) \quad D_y^2 u = P_2(y, x, D_x)u + P_1(y, x, D_x, D_y)u + f \quad \text{in } \Omega,$$

where

$$(1.2) \quad P_2(y, x, D_x) = \sum_{|\alpha|=2} a_\alpha(y, x) D_x^\alpha,$$

$$(1.3) \quad P_1(y, x, D_x, D_y) = \sum_{|\alpha| \leq 1} b_\alpha(y, x) D_x^\alpha + b(y, x) D_y.$$

We assume that all coefficients belong to $C^\infty(\Omega)$ and that $a_\alpha(y, x)$'s are real valued. Our main result in this section is the following theorem.

THEOREM 1.1. *Let $u \in D'(\Omega)$ be a solution of (1.1) in Ω and let $s \in \mathbb{R}$ be given. Suppose that $(y_0, x_0) \in \Omega$, $u \in H_{\text{loc}}^s(\Omega)$, and $f \in H_{\text{loc}}^{s-1}(\Omega)$. If we assume that*

$$(1.4) \quad (y_0, x_0, 0, \xi) \notin WF(u) \quad \text{for all } \xi \in \mathbb{R}^n \setminus \{0\}$$

and

$$(1.5) \quad f \text{ is microlocally } H^\mu \text{ at } (y_0, x_0, \pm 1, 0) \text{ for some } \mu > -\frac{3}{2},$$

then there is $\epsilon > 0$ and a neighborhood B of x_0 in \mathbb{R}^n such that

$$(1.6) \quad u \in C((y_0 - \epsilon, y_0 + \epsilon); H^s(B)).$$

If we further assume that

$$(1.7) \quad f \text{ is microlocally } H^\mu \text{ at } (y_0, x_0, \pm 1, 0) \text{ for some } \mu > -\frac{1}{2},$$

then

$$(1.8) \quad D_y u \in C((y_0 - \epsilon, y_0 + \epsilon); H^{s-1}(B)).$$

Remark 1.2. If $s > -\frac{1}{2}$, (1.5) is redundant and if $s > \frac{1}{2}$, (1.7) is redundant.

The remainder of this section is devoted to the proof of the above theorem. The strategy of the proof is to decompose u microlocally into three parts:

- (i) microlocally C^∞ part;
- (ii) elliptic part;
- (iii) hyperbolic part.

We then estimate each part separately. For this, we need some preparation. According to the above assumption (1.4), there are $\epsilon_1 > 0$, $0 < \delta_1 < \frac{1}{3}$ and a neighborhood B_1 of x_0 in \mathbb{R}^n such that

$$(1.9) \quad \Omega_1 \stackrel{\text{def}}{=} (y_0 - \epsilon_1, y_0 + \epsilon_1) \times B_1 \quad \text{and} \quad \overline{\Omega}_1 \subset \Omega,$$

$$(1.10) \quad \{\Omega_1 \times (\eta, \xi)\} \cap WF(u) = \emptyset, \quad \text{for all } (\eta, \xi) \in S^n \text{ satisfying } |\eta| \leq 3\delta_1.$$

Here $\overline{\Omega}_1$ is the closure of Ω_1 and S^n denotes the unit sphere in \mathbb{R}^{n+1} . Since we are interested in the behavior of u near (y_0, x_0) , we may assume that

$$(1.11) \quad u \in H^s(\mathbb{R}^{n+1}) \quad \text{with support in } \Omega,$$

$$(1.12) \quad f \in H^{s-1}(\mathbb{R}^{n+1}) \quad \text{with support in } \Omega,$$

and

$$(1.13) \quad D_y^2 u = P_2(y, x, D_x)u + P_1(y, x, D_x, D_y)u + f \quad \text{in } \Omega_1,$$

where we can also assume that

$$(1.14) \quad a_\alpha(y, x), b_\alpha(y, x), b(y, x) \in C_0^\infty(\mathbb{R}^{n+1}) \quad \text{with support in } \Omega.$$

Next we define

$$(1.15) \quad Q_1 = \{(y, x, \eta, \xi) : (y, x) \in \Omega_1, (\eta, \xi) \in S^n, |\eta| < 3\delta_1\},$$

$$(1.16) \quad Q_2 = \{(y, x, \eta, \xi) : (y, x) \in \Omega_1, (\eta, \xi) \in S^n, |\eta| > 2\delta_1, P_2(y, x, \xi) < \frac{1}{2}\eta^2\},$$

$$(1.17) \quad Q_3 = \{(y, x, \eta, \xi) : (y, x) \in \Omega_1, (\eta, \xi) \in S^n, |\eta| > 2\delta_1, P_2(y, x, \xi) > \frac{1}{3}\eta^2\}.$$

Then, $Q_1 \cup Q_2 \cup Q_3 = \Omega_1 \times S^n$.

We can find $q_i \in C_0^\infty(R^{n+1} \times S^n)$, $i = 1, 2, 3$, such that

$$(1.18) \quad \text{supp } q_i \subset Q_i, \quad i = 1, 2, 3,$$

$$(1.19) \quad 0 \leq q_i \leq 1, \quad i = 1, 2, 3$$

$$(1.20) \quad q_1 + q_2 + q_3 = 1 \quad \text{for all } (y, x, \eta, \xi) \in [y_0 - \epsilon_1/2, y_0 + \epsilon_1/2] \times \overline{B}_2 \times S^n,$$

where B_2 is a neighborhood of x_0 in R^n and its closure $\overline{B}_2 \subset B_1$.

Next let $\tilde{q}_i \in C^\infty(R^{n+1} \times R^{n+1})$ be an extension of q_i , $i = 1, 2, 3$, such that

$$(1.21) \quad \text{supp } \tilde{q}_i \subset \Omega_1 \times R^{n+1},$$

$$(1.22) \quad \tilde{q}_i = q_i \quad \text{on } R^{n+1} \times S^n,$$

$$(1.23) \quad \tilde{q}_i \text{ is homogeneous in } (\eta, \xi) \text{ of degree zero for } \eta^2 + |\xi|^2 \geq 1.$$

Then, it follows that $\tilde{q}_i \in S_{1,0}^0(R^{n+1})$. We can also choose a function $\rho_1(y, x) \in C_0^\infty(R^{n+1})$ such that

$$(1.24) \quad \text{supp } \rho_1 \subset (y_0 - \epsilon_1/2, y_0 + \epsilon_1/2) \times B_2,$$

$$(1.25) \quad \rho_1(y, x) = 1 \quad \text{on } [y_0 - \epsilon_1/3, y_0 + \epsilon_1/3] \times \overline{B}_3,$$

where B_3 is a neighborhood of x_0 in R^n and its closure $\overline{B}_3 \subset B_2$. Let us define $\Lambda_i \in OPS^0(R^{n+1})$, $i = 1, 2, 3$, by

$$(1.26) \quad \Lambda_i(v) = \sigma_i(\rho_1 v) \quad \text{for each } v \in \mathcal{D}'(R^{n+1}),$$

where $\sigma_i \in OPS^0(R^{n+1})$ has the symbol \tilde{q}_i .

Then, we find that

$$(1.27) \quad \rho_1 v = (\Lambda_1 + \Lambda_2 + \Lambda_3)v \text{ mod } C^\infty(R^{n+1})$$

for each $v \in \mathcal{D}'(R^{n+1})$, since

$$(1.28) \quad ES(I - (\sigma_1 + \sigma_2 + \sigma_3)) \cap \{(y_0 - \epsilon_1/2, y_0 + \epsilon_1/2) \times B_2 \times S^n\} = \phi,$$

where I stands for the identity mapping.

1.1. Regularity of $\Lambda_2 u$ for $s \leq -\frac{1}{2}$. We first note that $\text{supp } q_2$ is not empty by virtue of the definition of Q_i 's and (1.20). Since $(y_0, x_0, \pm 1, 0)$ does not belong to the characteristic set of the differential operator $D_y^2 - P_2(y, x, D_x) - P_1(y, x, D_x, D_y)$, u is microlocally $H^{\mu+2}$ at $(y_0, x_0, \pm 1, 0)$. Hence, there are $0 < \delta_2 < \frac{1}{3} - \delta_1$, a neighborhood Ω_2 of (y_0, x_0) in R^{n+1} , $q_4 \in C_0^\infty(R^{n+1} \times S^n)$, and $\Lambda_4 \in OPS^0(R^{n+1})$ such that

$$(1.29) \quad \overline{\Omega}_2 \subset \Omega_1,$$

$$(1.30) \quad \text{supp } q_4 \subset Q_4 \stackrel{\text{def}}{=} \Omega_1 \times \{(\eta, \xi) : (\eta, \xi) \in S^n, |\eta| > 1 - 3\delta_2\},$$

$$(1.31) \quad q_4 = 1 \quad \text{on } \overline{\Omega}_2 \times \{(\eta, \xi) : (\eta, \xi) \in S^n, |\eta| \geq 1 - 2\delta_2\},$$

$$(1.32) \quad \Lambda_4 u \in H^{\mu+2}(R^{n+1})$$

where the symbol of Λ_4 is $\tilde{q}_4(y, x, \eta, \xi) \in C^\infty(R^{n+1} \times R^{n+1})$ such that

$$(1.33) \quad \text{supp } \tilde{q}_4 \subset \Omega_1 \times R^{n+1},$$

$$(1.34) \quad \tilde{q}_4 = q_4 \quad \text{on } R^{n+1} \times S^n,$$

$$(1.35) \quad \tilde{q}_4 \text{ is homogeneous in } (\eta, \xi) \text{ of degree zero for } \eta^2 + |\xi|^2 \geq 1.$$

Next we choose $q_5 \in C_0^\infty(R^{n+1} \times S^n)$ with support in Q_2 such that

$$(1.36) \quad 0 \leq q_5 \leq 1$$

$$(1.37) \quad q_5 = 1 \quad \text{on an open subset of } \Omega_1 \times S^n \text{ which contains the support of } q_2.$$

Let $\Lambda_5 \in OPS^0(R^{n+1})$ have the symbol $\tilde{q}_5(y, x, \eta, \xi) \in C^\infty(R^{n+1} \times R^{n+1})$ which is an extension of q_5 satisfying the conditions analogous to (1.33), (1.34), and (1.35). We also choose $\rho_2 \in C_0^\infty(R^{n+1})$ such that

$$(1.38) \quad \text{supp } \rho_2 \subset \Omega_2,$$

$$(1.39) \quad \rho_2 = 1 \quad \text{on } \overline{\Omega}_3 \subset \Omega_2,$$

where Ω_3 is a neighborhood of (y_0, x_0) . Then it follows that

$$(1.40) \quad (I - \Lambda_5)\rho_2\Lambda_2 \in OPS^{-\infty}(R^{n+1})$$

and consequently,

$$(1.41) \quad \left(D_y^2 + \sum_{i=1}^n D_{x_i}^2 \right) (I - \Lambda_5)\rho_2\Lambda_2 \in OPS^{-\infty}(R^{n+1}).$$

Let us define

$$(1.42) \quad \mathcal{L} = D_y^2\Lambda_5 - P_2(y, x, D_x)\Lambda_5 + \left(D_y^2 + \sum_{i=1}^n D_{x_i}^2 \right) (I - \Lambda_5).$$

Then, \mathcal{L} is elliptic and it follows from (1.13) that

$$(1.43) \quad \mathcal{L}\rho_2\Lambda_2u = h_1 \bmod C_0^\infty(R^{n+1}),$$

where

$$(1.44) \quad h_1 = \Lambda_5\rho_2\Lambda_2P_1(y, x, D_x, D_y)u + \Lambda_5\rho_2\Lambda_2f + [D_y^2 - P_2(y, x, D_x), \Lambda_5\rho_2\Lambda_2]u,$$

and we used $(I - \Lambda_5)\rho_2\Lambda_2u \in C_0^\infty(R^{n+1})$.

Hence, we have

$$(1.45) \quad \mathcal{L}\rho_2\Lambda_2(I - \Lambda_4)u = h_2 \bmod C_0^\infty(R^{n+1}),$$

$$(1.46) \quad h_2 = [\mathcal{L}\rho_2\Lambda_2, I - \Lambda_4]u + (I - \Lambda_4)h_1.$$

Now we observe that

$$(1.47) \quad h_1, h_2 \in H^{s-1}(R^{n+1}),$$

$$(1.48) \quad WF(h_1) \subset ES(\Lambda_2) \cap (\Omega_2 \times R^{n+1}),$$

$$(1.49) \quad \begin{aligned} WF(h_2) \subset & \{ES(I - \Lambda_4) \cap ES(\Lambda_2) \cap (\Omega_2 \times R^{n+1})\} \\ & \cup \{ES(I - \Lambda_4) \cap WF(h_1)\}, \end{aligned}$$

from which we can deduce that

$$(1.50) \quad WF(h_2) \cap [R^{n+1} \times \{(\eta, \xi) : (\eta, \xi) \in S^n, |\eta| < 2\delta_1 \text{ or } |\eta| > 1 - 2\delta_2\}] = \phi.$$

For (1.48) and (1.49), the reader may refer to [19, pp. 127–128]. Next let \mathcal{M} be a parametrix for \mathcal{L} . Then (1.45) yields

$$(1.51) \quad \rho_2\Lambda_2(I - \Lambda_4)u = \mathcal{M}h_2 \bmod C^\infty(R^{n+1}).$$

Finally we choose $q_6(\eta, \xi) \in C^\infty(S^n)$ such that

$$(1.52) \quad \text{supp } q_6 \subset \{(\eta, \xi) : (\eta, \xi) \in S^n, \delta_1 < |\eta| < 1 - \delta_2\}$$

$$(1.53) \quad q_6 = 1 \quad \text{for } \frac{3}{2}\delta_1 \leq |\eta| \leq 1 - \frac{3}{2}\delta_2, \quad (\eta, \xi) \in S^n.$$

Let $\Lambda_6 \in OPS^0(R^{n+1})$ have the symbol $\tilde{q}_6(\eta, \xi) \in C^\infty(R^{n+1})$ which coincides with q_6 on S^n and is homogeneous in (η, ξ) of degree zero for $\eta^2 + |\xi|^2 \geq 1$. Then, we find that

$$(1.54) \quad WF((I - \Lambda_6)\mathcal{M}h_2) \subset ES(I - \Lambda_6) \cap WF(h_2)$$

and this is empty by virtue of (1.50) and (1.53). Consequently,

$$(1.55) \quad \mathcal{M}h_2 = \Lambda_6\mathcal{M}h_2 \bmod C^\infty(R^{n+1}).$$

Let us set

$$(1.56) \quad h_3 = \mathcal{M}h_2.$$

Then, $h_3 \in H^{s+1}(R^{n+1})$ since $\mathcal{M} \in OPS^{-2}(R^{n+1})$. We also find that

$$(1.57) \quad \mathcal{F}(\Lambda_6 h_3)(\eta, \xi) = \tilde{q}_6(\eta, \xi) \mathcal{F}(h_3)(\eta, \xi)$$

where \mathcal{F} denotes the Fourier transform.

By virtue of (1.52), we see that for all (η, ξ) satisfying $\eta^2 + |\xi|^2 \geq 1$ and $k = 1, 2$,

$$(1.58) \quad \begin{aligned} & (1 + |\eta|)^k (1 + |\xi|)^{s+1-k} |\tilde{q}_6(\eta, \xi) \mathcal{F}(h_3)(\eta, \xi)| \\ & \leq c(1 + |\eta| + |\xi|)^{s+1} |\tilde{q}_6(\eta, \xi) \mathcal{F}(h_3)(\eta, \xi)| \end{aligned}$$

holds for some positive constant c . Since $h_3 \in H^{s+1}(R^{n+1})$, it follows from (1.58) that

$$(1.59) \quad \Lambda_6 h_3 \in C(R; H^s(R^n)) \cap C^1(R; H^{s-1}(R^n)).$$

By virtue of (1.51), (1.55), and (1.56), we find that

$$(1.60) \quad \rho_2 \Lambda_2 (I - \Lambda_4) u \in C(R; H_{\text{loc}}^s(R^n)) \cap C^1(R; H_{\text{loc}}^{s-1}(R^n)),$$

and thus, on account of (1.38),

$$(1.61) \quad \rho_2 \Lambda_2 (I - \Lambda_4) u \in C(R; H^s(R^n)) \cap C^1(R; H^{s-1}(R^n)).$$

Since $\Lambda_2 \in OPS^0(R^{n+1})$, (1.32) implies that

$$(1.62) \quad \Lambda_2 \Lambda_4 u \in C(R; H^{\mu+3/2}(R^n)) \quad \text{if } \mu > -\frac{3}{2}$$

and

$$(1.63) \quad \Lambda_2 \Lambda_4 u \in C(R; H^{\mu+3/2}(R^n)) \cap C^1(R; H^{\mu+1/2}(R^n)) \quad \text{if } \mu > -\frac{1}{2}.$$

Combining (1.61), (1.62), and (1.63), we conclude that

$$(1.64) \quad \rho_2 \Lambda_2 u \in C(R; H^s(R^n)) \quad \text{if } \mu > -\frac{3}{2},$$

$$(1.65) \quad \rho_2 \Lambda_2 u \in C(R; H^s(R^n)) \cap C^1(R; H^{s-1}(R^n)) \quad \text{if } \mu > -\frac{1}{2}.$$

1.2. Regularity of $\Lambda_2 u$ for $s > -\frac{1}{2}$. For $-\frac{1}{2} < s \leq \frac{1}{2}$, we take $\mu = s - 1$ to obtain (1.64), and obtain (1.65) under assumption (1.7). For $s > \frac{1}{2}$, we simply take $\mu = s - 1$ to have (1.65).

1.3. Regularity of $\Lambda_3 u$. If $\text{supp } q_3$ is empty, then Λ_3 is vacuous and we skip over this section. So we assume that $\text{supp } q_3$ is not empty in this section. We define

$$(1.66) \quad G_1 = \{(y, x, \xi) : (y, x) \in \Omega_1, \xi \in S^{n-1}, P_2(y, x, \xi) > \delta_1^2 / (1 - 4\delta_1^2)\},$$

$$(1.67) \quad \Omega_3 = \{(y, x) : (y, x) \in \Omega_1, q_3(y, x, \eta, \xi) \neq 0 \text{ for some } (\eta, \xi) \in S^n\},$$

$$(1.68) \quad G_2 = \{(y, x, \xi) : (y, x) \in \Omega_3, \xi \in S^{n-1}, P_2(y, x, \xi) > 4\delta_1^2 / 3(1 - 4\delta_1^2)\},$$

where Ω_1 and δ_1 were defined in the above section and S^{n-1} denotes the unit sphere in R^n . Since $\text{supp } q_3$ is not empty, Ω_3 is not empty. We note that if $(y, x, \eta, \xi) \in Q_3$, then $\xi \neq 0$, $\eta^2 + |\xi|^2 = 1$, $|\eta| > 2\delta_1$ and

$$(1.69) \quad P_2(y, x, \xi) > \frac{1}{3}\eta^2 > 4\delta_1^2 |\xi|^2 / 3(1 - 4\delta_1^2),$$

since $1 - 4\delta_1^2 > |\xi|^2$. Therefore, G_2 is not empty. On account of (1.18), $\overline{\Omega}_3 \subset \Omega_1$ and thus, $\overline{G}_2 \subset G_1$. Next we choose $\psi \in C_0^\infty(R^{n+1} \times S^{n-1})$ such that

$$(1.70) \quad \text{supp } \psi \subset G_1,$$

$$(1.71) \quad 0 \leq \psi \leq 1,$$

$$(1.72) \quad \psi = 1 \quad \text{on } \overline{G}_2.$$

Let $\tilde{\psi} \in C^\infty(R^{n+1} \times R^n)$ be an extension of ψ such that

$$(1.73) \quad \text{supp } \tilde{\psi} \subset \Omega_1 \times R^n,$$

$$(1.74) \quad \tilde{\psi} = \psi \quad \text{on } R^{n+1} \times S^{n-1},$$

$$(1.75) \quad \tilde{\psi} \text{ is homogeneous in } \xi \text{ of degree zero for } |\xi| \geq 1.$$

Let $\Psi \in OPS^0(R^n)$ have the symbol $\tilde{\psi}$ so that Ψ depends smoothly on $y \in R$.

LEMMA 1.3. *It holds that*

$$(1.76) \quad \Psi \Lambda_3 \in OPS^0(R^{n+1}),$$

$$(1.77) \quad \text{supp } \Psi \Lambda_3 v \subset \Omega_1 \quad \text{for every } v \in \mathcal{D}'(R^{n+1}),$$

$$(1.78) \quad (I - \Psi) \Lambda_3 \in OPS^{-\infty}(R^{n+1}).$$

Proof. Since $\text{supp } q_3 \subset Q_3$, there is a positive constant c such that

$$(1.79) \quad \tilde{q}_3(y, x, \eta, \xi) = 0 \quad \text{if } |\xi| \leq c|\eta| \quad \text{and} \quad \eta^2 + |\xi|^2 \geq 1.$$

Recalling that Λ_3 was defined by (1.26), we can use Theorem 18.1.35 of [6] with the help of (1.79) to derive (1.76). For (1.77), we choose any $v_1 \in C_0^\infty(R^{n+1})$ and set

$$(1.80) \quad v_2 = \Lambda_3 v_1.$$

Then, $v_2 \in C_0^\infty(R^{n+1})$ with support in Ω_1 and

$$(1.81) \quad \Psi v_2 = (2\pi)^{-n} \int_{R^n} \tilde{\psi}(y, x, \xi) \exp(ix \cdot \xi) \mathcal{F}_x(v_2)(y, \xi) d\xi,$$

where

$$(1.82) \quad \mathcal{F}_x(v_2)(y, \xi) = \int_{R^n} \exp(-ix \cdot \xi) v_2(y, x) dx.$$

It follows from (1.73) and (1.81) that

$$(1.83) \quad \text{supp } \Psi v_2 \subset \Omega_1.$$

This implies (1.77). For (1.78), we use (1.68), (1.69), (1.72), and (1.79) to see that

$$(1.84) \quad \begin{aligned} \text{supp } \tilde{q}_3 \cap \{ & (y, x, \eta, \xi) : (y, x) \in R^{n+1}, \eta^2 + |\xi|^2 \geq 1 \} \\ & \subset \{ (y, x, \eta, \xi) : (y, x) \in \Omega_1, |\xi| \geq 1, |\xi| \geq c|\eta| \text{ and } \tilde{\psi}(y, x, \xi) = 1 \} \\ & \cup \{ (y, x, \eta, \xi) : (y, x) \in \Omega_1, 1 \geq |\xi| \geq c|\eta| \}, \end{aligned}$$

where the positive constant c is the same size as in (1.79). Now (1.78) follows from (1.84). This ends the proof of the lemma.

Next we derive from (1.13) that

$$(1.85) \quad \begin{aligned} (D_y^2 - P_2(y, x, D_x))\Psi\Lambda_3 u &= \Psi\Lambda_3 P_1(y, x, D_x, D_y)u \\ &+ \Psi\Lambda_3 f + [D_y^2 - P_2(y, x, D_x), \Psi\Lambda_3]u, \end{aligned}$$

and define

$$(1.86) \quad A = P_2(y, x, D_x)\Psi + \sum_{i=1}^n D_{x_i}^2(I - \Psi).$$

Then, A is elliptic in R^n with y as a parameter. By virtue of (1.78), it is evident that

$$(1.87) \quad (D_y^2 - P_2(y, x, D_x))\Psi\Lambda_3 u = (D_y^2 - A)\Lambda_3 u \bmod C^\infty(R^{n+1}),$$

and consequently,

$$(1.88) \quad \begin{aligned} (D_y^2 - A)\Lambda_3 u &= \Psi\Lambda_3 P_1(y, x, D_x, D_y)u \\ &+ \Psi\Lambda_3 f + [D_y^2 - P_2(y, x, D_x), \Psi\Lambda_3]u \bmod C^\infty(R^{n+1}). \end{aligned}$$

We set

$$(1.89) \quad w = \Lambda_3 u$$

and note that

$$(1.90) \quad w \in H^s(R^{n+1}),$$

$$(1.91) \quad \text{supp } w \subset \Omega_1,$$

$$(1.92) \quad (D_y^2 - A)w = h_4 \bmod C^\infty(R^{n+1}),$$

where

$$(1.93) \quad \begin{aligned} h_4 &\stackrel{\text{def}}{=} \Psi\Lambda_3 P_1(y, x, D_x, D_y)u + \Psi\Lambda_3 f \\ &+ [D_y^2 - P_2(y, x, D_x), \Psi\Lambda_3]u. \end{aligned}$$

By means of (1.76) and (1.77), it is easy to see that

$$(1.94) \quad h_4 \in H^{s-1}(R^{n+1}),$$

$$(1.95) \quad \text{supp } h_4 \subset \Omega_1.$$

Since $WF(w) \subset ES(\Lambda_3)$ and $WF(h_4) \subset ES(\Lambda_3)$, it follows that

$$(1.96) \quad WF(w) \cap (R^{n+1} \times S^n) \subset Q_3,$$

$$(1.97) \quad WF(h_4) \cap (R^{n+1} \times S^n) \subset Q_3.$$

Next we choose positive constants c_1 and c_2 such that

$$(1.98) \quad Q_3 \subset \Omega_1 \times \{(\eta, \xi) : (\eta, \xi) \in S^n, c_1|\eta| \leq |\xi| \leq c_2|\eta|\},$$

and let $\Lambda_7 \in OPS^0(R^{n+1})$ have the symbol $q_7(\eta, \xi) \in C^\infty(R^{n+1})$ such that

$$(1.99) \quad \text{supp } q_7 \cap S^n \subset \{(\eta, \xi) : (\eta, \xi) \in S^n, \tfrac{1}{2}c_1|\eta| \leq |\xi| \leq 2c_2|\eta|\},$$

$$(1.100) \quad q_7 = 1 \quad \text{on } \{(\eta, \xi) : (\eta, \xi) \in S^n, c_1|\eta| \leq |\xi| \leq c_2|\eta|\}$$

$$(1.101) \quad q_7 \text{ is homogeneous in } (\eta, \xi) \text{ of degree zero for } \eta^2 + |\xi|^2 \geq 1.$$

Then, by means of (1.96)–(1.101), we can infer that

$$(1.102) \quad w = \Lambda_7 w \bmod C^\infty(R^{n+1})$$

$$(1.103) \quad h_4 = \Lambda_7 h_4 \bmod C^\infty(R^{n+1}).$$

Since the Fourier transform of $\Lambda_7 w$ is given by

$$(1.104) \quad \mathcal{F}(\Lambda_7 w)(\eta, \xi) = q_7(\eta, \xi) \mathcal{F}(w)(\eta, \xi),$$

we derive from (1.90), (1.99), and (1.101) that

$$(1.105) \quad \Lambda_7 w \in H^3(R; H^{s-3}(R^n))$$

which, together with (1.91) and (1.102), yields

$$(1.106) \quad w \in C^2(R; H^{s-3}(R^n)).$$

Similarly, by virtue of (1.94), (1.99), and (1.101), we find that

$$(1.107) \quad \Lambda_7 h_4 \in L^2(R; H^{s-1}(R^n))$$

and thus, by (1.95) and (1.103),

$$(1.108) \quad h_4 \in L^2(R; H^{s-1}(R^n)).$$

In the meantime, it follows from (1.77), (1.86), and (1.89) that

$$(1.109) \quad \text{supp } Aw \subset \Omega_1.$$

Consequently, we can rewrite (1.92), with the help of (1.95) and (1.109), as

$$(1.110) \quad (D_y^2 - A)w = h_4 + h_5,$$

where $h_5 \in C_0^\infty(R^{n+1})$.

According to Theorem 23.2.2 of [6], there is a unique solution

$$(1.111) \quad \tilde{w} \in C([y_0 - \epsilon_1, y_0 + \epsilon_1]; H^s(R^n)) \cap C^1([y_0 - \epsilon_1, y_0 + \epsilon_1]; H^{s-1}(R^n))$$

of

$$(1.112) \quad (D_y^2 - A)\tilde{w} = h_4 + h_5,$$

$$(1.113) \quad \tilde{w} = 0, \quad D_y \tilde{w} = 0 \quad \text{for } y = y_0 - \epsilon_1.$$

Again by Theorem 23.2.2 of [6], the uniqueness of solution is still valid for lower s . Thus, $w = \tilde{w}$ and

$$(1.114) \quad \Lambda_3 u \in C([y_0 - \epsilon_1, y_0 + \epsilon_1]; H^s(R^n)) \cap C^1([y_0 - \epsilon_1, y_0 + \epsilon_1]; H^{s-1}(R^n)).$$

1.4. Conclusion of the proof. We derive from (1.10), (1.15), and (1.21) that

$$(1.115) \quad \Lambda_1 u \in C_0^\infty(R^{n+1})$$

with support in Ω_1 .

We combine (1.27), (1.64), (1.65), (1.114), and (1.115) to arrive at (1.6) and (1.8).

2. Application to exact boundary control of a wave equation. As mentioned earlier, Russell [16], [17] obtained the exact boundary control of a wave equation via Huygen's Principle. This idea was further developed by Lagnese [8], [9] and Littman [14], [15]. Our purpose in this section is to improve the regularity of the boundary control obtained in [8] and [16]. We will discuss each method separately.

2.1. Russell's method. We consider a wave equation

$$(2.1) \quad \frac{\partial^2 u}{\partial t^2} - \Delta u = 0 \quad \text{in } \mathcal{O} \times (0, T),$$

$$(2.2) \quad u(z, 0) = u_0(z), \quad \frac{\partial u}{\partial t}(z, 0) = v_0(z) \quad \text{in } \mathcal{O},$$

$$(2.3) \quad \alpha u(z, t) + \beta \frac{\partial u}{\partial \nu}(z, t) = F(z, t) \quad \text{on } \partial \mathcal{O} \times (0, T).$$

Here, \mathcal{O} is a bounded domain in R^n with smooth boundary $\partial \mathcal{O}$, $\partial/\partial \nu$ denotes the outward normal derivative on $\partial \mathcal{O}$, and α, β are nonnegative constants such that $\alpha + \beta > 0$. The issue of exact boundary controllability is to find $F(z, t)$ for a given $u_0(z)$ and $v_0(z)$ such that the solution of (2.1), (2.2), and (2.3) satisfies

$$(2.4) \quad u(z, T) = 0, \quad \frac{\partial u}{\partial t}(z, T) = 0 \quad \text{in } \mathcal{O}.$$

We will sketch Russell's method when $n \geq 3$ is odd. Let \mathcal{O}_δ be an open bounded domain in R^n such that $\overline{\mathcal{O}} \subset \mathcal{O}_\delta$ and distance $(\partial \mathcal{O}, \partial \mathcal{O}_\delta) < \delta$. Let $(u_0, v_0) \in H^2(\mathcal{O}) \times H^1(\mathcal{O})$ be given. We can extend (u_0, v_0) to $(u_\delta, v_\delta) \in H^2(R^n) \times H^1(R^n)$ such that $u_\delta(z) = v_\delta(z) = 0$ for $z \notin \mathcal{O}_\delta$. Then, we solve the Cauchy problem for (2.1) in $R^n \times (0, \infty)$ with initial data (u_δ, v_δ) . Then, the solution w satisfies

$$(2.5) \quad w(z, t) = 0 \quad \text{for all } z \in \mathcal{O} \quad \text{and} \quad t > 2\delta + \text{diameter of } \mathcal{O},$$

which follows from Huygen's principle. We set

$$(2.6) \quad F(z, t) = \alpha w(z, t) + \beta \frac{\partial w}{\partial \nu}(z, t)$$

for $(z, t) \in \partial \mathcal{O} \times (0, T)$, where $T > 2\delta + \text{diameter of } \mathcal{O}$, and find the solution u of (2.1), (2.2), and (2.3). By the uniqueness of solution to the initial-boundary value problem (2.1), (2.2), and (2.3), $u = w$ in $\mathcal{O} \times (0, T)$ and consequently, $F(z, t)$ given by (2.6) is the desired control.

THEOREM 2.1. *If \mathcal{O} is convex and $(u_0, v_0) \in H^2(\mathcal{O}) \times H^1(\mathcal{O})$ has compact support in \mathcal{O} , then the above control given by (2.6) has the following regularity:*

$$(2.7) \quad F \in H^1(\partial \mathcal{O} \times (0, T)) \quad \text{if } \beta \neq 0,$$

$$(2.8) \quad F \in H^2(\partial \mathcal{O} \times (0, T)) \quad \text{if } \beta = 0.$$

This follows from the following discussion of a more general situation. For given $(u_0, v_0) \in H^s(R^n) \times H^{s-1}(R^n)$, $s \in R$, let $u \in C(R; H^s(R^n)) \cap C^1(R; H^{s-1}(R^n))$ be the unique solution of the following Cauchy problem:

$$(2.9) \quad \frac{\partial^2 u}{\partial t^2} - \Delta u = 0 \quad \text{in } R^n \times R,$$

$$(2.10) \quad u(z, 0) = u_0(z), \quad \frac{\partial u}{\partial t}(z, 0) = v_0(z) \quad \text{in } R^n.$$

and set

$$(2.11) \quad g(z, t) = \text{the restriction of } u(z, t) \quad \text{to } \partial\mathcal{O} \times R,$$

$$(2.12) \quad h(z, t) = \text{the restriction of } \frac{\partial u}{\partial \nu}(z, t) \quad \text{to } \partial\mathcal{O} \times R.$$

THEOREM 2.2. *Suppose that \mathcal{O} is convex and that $\text{supp } u_0 \cup \text{supp } v_0 \subset K$, for some compact subset $K \subset \mathcal{O}$. Then, for any $T > 0$, it holds that*

$$(2.13) \quad g \in H^s(\partial\mathcal{O} \times (-T, T)),$$

$$(2.14) \quad h \in H^{s-1}(\partial\mathcal{O} \times (-T, T)),$$

$$(2.15) \quad \|g\|_{H^s(\partial\mathcal{O} \times (-T, T))} + \|h\|_{H^{s-1}(\partial\mathcal{O} \times (-T, T))} \leq C(\|u_0\|_{H^s(R^n)} + \|v_0\|_{H^{s-1}(R^n)})$$

for some positive constant C which depends on K and T , but is independent of u_0 and v_0 .

Proof. Choose any $(z_0, t_0) \in \partial\mathcal{O} \times (-\infty, \infty)$. Each bicharacteristic strip of (2.9) is a straight line in $R^{n+1} \times R^{n+1}$. We call its projection onto the base space R^{n+1} a bicharacteristic curve. Since \mathcal{O} is convex and $\text{supp } u_0 \cup \text{supp } v_0 \subset K$, each bicharacteristic curve which passes through (z_0, t_0) and is tangent to the surface $\partial\mathcal{O} \times R$ at (z_0, t_0) does not meet $\text{supp } u_0 \cup \text{supp } v_0$ at $t = 0$. Since the waves propagate at finite speed, it is easily seen that $(z, 0, \zeta, \tau) \notin WF(u)$ for any $(\zeta, \tau) \in R^{n+1} \setminus \{0\}$ if $z \notin \text{supp } u_0 \cup \text{supp } v_0$. Thus, by virtue of Proposition 3.5.1 of [5], we find that

$$(2.16) \quad (z_0, t_0, \zeta, \tau) \notin WF(u)$$

if $(\zeta, \tau) \in R^{n+1} \setminus \{0\}$ is tangent to $\partial\mathcal{O} \times R$ at (z_0, t_0) . Furthermore, according to Corollary C.5.3 of [6], there is a local coordinate transformation from (z, t) to (y, x) such that:

- (i) (z_0, t_0) is mapped to (y_0, x_0) ;
- (ii) t variable is not affected and becomes one component of the x variables;
- (iii) the points of $\partial\mathcal{O} \times R$ near (z_0, t_0) are characterized by $y = y_0$;
- (iv) the points of $\mathcal{O} \times R$ near (z_0, t_0) are characterized by $y > y_0$;
- (v) equation (2.9) is transformed into (1.1) with $f = 0$, which is valid near (y_0, x_0) .

This coordinate transformation induces a local diffeomorphism of the cotangent bundle. Under this diffeomorphism, we find that (z_0, t_0, ζ, τ) is mapped to $(y_0, x_0, 0, \xi)$ for some $\xi \in R^n \setminus \{0\}$ if and only if $(\zeta, \tau) \in R^{n+1} \setminus \{0\}$ is tangent to the surface $\partial\mathcal{O} \times R$ at (z_0, t_0) .

In the meantime, the wave front set is invariant under the diffeomorphism of coordinates. Therefore, we conclude that

$$(2.17) \quad (y_0, x_0, 0, \xi) \notin WF(\tilde{u}),$$

for each $\xi \in R^n \setminus \{0\}$, where we write

$$(2.18) \quad \tilde{u}(y, x) = u(z(y, x), t(y, x)).$$

Now (2.13) and (2.14) follow from (1.6) and (1.8). It remains to prove (2.15). In the above setting, we let \mathcal{T} be a linear operator that maps $(u_0(z), v_0(z))$ to $\tilde{u}(y, x)$, which was defined by (2.18). Next we write

$$(2.19) \quad H_K^s(R^n) = \{w \in H^s(R^n) : \text{supp } w \subset K\}.$$

If $u(z, t)$ is a solution of (2.9) and (2.10) with $(u_0, v_0) \in H_K^s(R^n) \times H_K^{s-1}(R^n)$, then for each $(\zeta, \tau) \in R^{n+1} \setminus \{0\}$ which is tangent to $\partial\mathcal{O} \times R$ at (z_0, t_0) , we can find a conic neighborhood of (z_0, t_0, ζ, τ) in $R^{n+1} \times (R^{n+1} \setminus \{0\})$ which is disjoint from $WF(u)$. We can choose this conic neighborhood independently of (u_0, v_0) since K is fixed. Consequently, ϵ_1, B_1 , and δ_1 in (1.9) and (1.10) can be taken independently of (u_0, v_0) . Since f in (1.1) vanishes in this section, δ_2 in (1.30) and Λ_4 in (1.32) can be taken independently of (u_0, v_0) . Hence, ϵ and B in (1.6) and (1.8) are independent of (u_0, v_0) .

Now we conclude that \mathcal{T} is a linear mapping from $H_K^s(R^n) \times H_K^{s-1}(R^n)$ into $C((-\epsilon, \epsilon); H^s(B)) \cap C^1((-\epsilon, \epsilon); H^{s-1}(B))$. It is apparent that \mathcal{T} is a closed operator and thus, \mathcal{T} is continuous. It follows that there is a neighborhood Σ of (z_0, t_0) in $\partial\mathcal{O} \times R$ such that

$$(2.20) \quad \|g\|_{H^s(\Sigma)} + \|h\|_{H^{s-1}(\Sigma)} \leq C(\|u_0\|_{H_K^s(R^n)} + \|v_0\|_{H_K^{s-1}(R^n)})$$

for some constant C independent of u_0 and v_0 . Since $\partial\mathcal{O} \times [-T, T]$ can be covered by a finite number of such neighborhoods, we arrive at (2.15).

Next we will extend Theorem 2.1 to the case where (u_0, v_0) is less regular. We need to define a weak solution to the initial-boundary value problem. For this, we first define

$$(2.21) \quad \tilde{H}_\rho^s(\partial\mathcal{O} \times (0, T)) = \{g \in H^s(\partial\mathcal{O} \times (0, T)) : g \text{ vanishes for } 0 < t < \rho\},$$

which is a closed subspace of $H^s(\partial\mathcal{O} \times (0, T))$. As above, we assume that K is a compact subset of \mathcal{O} .

DEFINITION 2.3. Let $s \in R$. Suppose that $(u_0, v_0) \in H_K^s(R^n) \times H_K^{s-1}(R^n)$ and $g \in \tilde{H}_\rho^s(\partial\mathcal{O} \times (0, T))$. Then, a function $u(z, t) \in C([0, T]; H^s(\mathcal{O})) \cap C^1([0, T]; H^{s-1}(\mathcal{O}))$ is called a solution of the initial-boundary value problem (2.1), (2.2), and

$$(2.22) \quad u(z, t) = g(z, t) \quad \text{on } \partial\mathcal{O} \times (0, T)$$

if u satisfies

$$(2.23) \quad \int_0^T \langle u(z, t), \chi(z, t) \rangle_1 dt = \langle v_0(z), \theta(z, 0) \rangle_2 - \left\langle u_0(z), \frac{\partial \theta}{\partial t}(z, 0) \right\rangle_3 - \left\langle g(z, t), \frac{\partial \theta}{\partial \nu}(z, t) \right\rangle_4$$

for any $\chi \in C_0^\infty(\mathcal{O} \times (0, T))$, where $\theta(z, t) \in C^\infty(\overline{\mathcal{O}} \times [0, T])$ is the solution of

$$(2.24) \quad \frac{\partial^2 \theta}{\partial t^2} - \Delta \theta = \chi \quad \text{in } \mathcal{O} \times (0, T),$$

$$(2.25) \quad \theta(z, T) = 0, \quad \frac{\partial \theta}{\partial t}(z, T) = 0 \quad \text{in } \mathcal{O},$$

$$(2.26) \quad \theta(z, t) = 0 \quad \text{on } \partial \mathcal{O} \times (0, T),$$

and we use the following notation:

$$(2.27) \quad \langle \cdot, \cdot \rangle_1 = \text{the duality pairing between } H^s(\mathcal{O}) \text{ and its dual,}$$

$$(2.28) \quad \langle \cdot, \cdot \rangle_2 = \text{the duality pairing between } H_c^{s-1}(\mathcal{O}) \text{ and its dual } H_{\text{loc}}^{1-s}(\mathcal{O}),$$

$$(2.29) \quad \langle \cdot, \cdot \rangle_3 = \text{the duality pairing between } H_c^s(\mathcal{O}) \text{ and its dual } H_{\text{loc}}^{-s}(\mathcal{O}),$$

$$(2.30) \quad \langle \cdot, \cdot \rangle_4 = \text{the duality pairing between } \tilde{H}_\rho^s(\partial \mathcal{O} \times (0, T)) \text{ and its dual.}$$

Here ρ is a positive number such that $\rho < \frac{1}{2}$ distance $(\partial \mathcal{O}, K)$. Each bracket reduces to the L^2 inner product when both elements belong to L^2 . Since $\chi \in C_0^\infty(\mathcal{O} \times (0, T))$, (2.25) and (2.26) imply that θ vanishes near $t = T$. Hence, the last term of (2.23) is well defined.

PROPOSITION 2.4. *Let $u(z, t)$ be the solution of the Cauchy problem (2.9) and (2.10) with $(u_0, v_0) \in H_K^s(R^n) \times H_K^{s-1}(R^n)$. Let g be defined by (2.11). Then, for this (u_0, v_0) and g , $u(z, t)$ is a unique solution of the initial-boundary value problem (2.1), (2.2), and (2.22).*

Proof. Fix any $\chi(z, t) \in C_0^\infty(\mathcal{O} \times (0, T))$ and let θ be a solution of (2.24), (2.25), and (2.26). Choose a sequence $\{(u_0^k, v_0^k)\}_{k=1}^\infty$ in $C_0^\infty(R^n) \times C_0^\infty(R^n)$ with support in some compact subset K_1 such that (u_0^k, v_0^k) converges to (u_0, v_0) in $H_K^s(R^n) \times H_K^{s-1}(R^n)$ where $K \subset K_1 \subset \mathcal{O}$ and distance $(\partial \mathcal{O}, K_1) > \frac{1}{2}$ distance $(\partial \mathcal{O}, K)$. Let $u^k(z, t)$ be a solution of (2.9) with initial data (u_0^k, v_0^k) . Then, it is evident that u^k satisfies

$$(2.31) \quad \int_0^T \langle u^k(z, t), \chi(z, t) \rangle_1 dt = \langle v_0^k(z), \theta(z, 0) \rangle_2 - \left\langle u_0^k(z), \frac{\partial \theta}{\partial t}(z, 0) \right\rangle_3 \\ - \left\langle g^k(z, t), \frac{\partial \theta}{\partial \nu}(z, t) \right\rangle_4,$$

where g^k is the restriction of u^k to $\partial \mathcal{O} \times R$. We note that $g^k \in \tilde{H}_\rho^s(\partial \mathcal{O} \times (0, T))$ for each k since distance $(\partial \mathcal{O}, K_1) > \frac{1}{2}$ distance $(\partial \mathcal{O}, K) > \rho$. By virtue of (2.15), we pass $k \rightarrow \infty$ to see that u satisfies (2.23). The uniqueness is trivial.

Now we can assert the following theorem.

THEOREM 2.5. *Let $T > \text{diameter of } \mathcal{O}$. If \mathcal{O} is convex and $(u_0, v_0) \in H^s(\mathcal{O}) \times H^{s-1}(\mathcal{O})$, $s \in R$ has compact support in \mathcal{O} , then there is a control $F \in H^s(\partial \mathcal{O} \times (0, T))$ which drives the solution of (2.1), (2.2), and (2.3) with $\beta = 0$ to (2.4).*

By modifying Definition 2.3 in an obvious manner, we can also define a weak solution to the initial-boundary value problem with the Neumann boundary condition. By repeating the same procedure as above, we can assert the following theorem.

THEOREM 2.6. *Under the same assumption as in Theorem 2.5, there is a control $F \in H^{s-1}(\partial \mathcal{O} \times (0, T))$ which drives the solution of (2.1), (2.2), and (2.3) with $\alpha = 0$ to (2.4).*

2.2. Lagnese's method. Lagnese [8] discussed a more general equation than a wave equation. But we will consider only a wave equation when the space dimension is even. In order to obtain a more regular boundary control, we will slightly modify the procedure.

For a given open bounded domain \mathcal{O} with smooth boundary, we define open bounded domains \mathcal{O}_δ and $\mathcal{O}_{2\delta}$ with smooth boundary such that

$$(2.32) \quad \overline{\mathcal{O}} \subset \mathcal{O}_\delta \subset \overline{\mathcal{O}_\delta} \subset \mathcal{O}_{2\delta},$$

$$(2.33) \quad \text{distance}(\partial\mathcal{O}, \partial\mathcal{O}_\delta) < \delta,$$

$$(2.34) \quad \text{distance}(\partial\mathcal{O}_\delta, \partial\mathcal{O}_{2\delta}) < \delta,$$

where δ is a small positive number. We employ a total extension operator E for \mathcal{O}_δ which maps $H^m(\mathcal{O}_\delta)$ into $H^m(R^n)$ continuously for each $m \geq 0$ such that for $h \in H^m(\mathcal{O}_\delta)$,

$$(2.35) \quad (Eh)(x) = h(x) \quad \text{for } x \in \mathcal{O}_\delta,$$

$$(2.36) \quad (Eh)(x) = 0 \quad \text{for } x \notin \mathcal{O}_{2\delta}.$$

For details, see [1]. We note that this total extension operator was not necessary in [8]. We suppose that

$$(2.37) \quad s \geq 2,$$

$$(2.38) \quad T_0 > 2\delta + \text{diameter of } \mathcal{O}_\delta$$

and construct a linear operator Λ_T , for $T \geq T_0$ as follows. For given $(u_0, v_0) \in H^s(\mathcal{O}_\delta) \times H^{s-1}(\mathcal{O}_\delta)$, let $w \in C([0, \infty); H^s(R^n)) \cap C^1([0, \infty); H^{s-1}(R^n))$ be the solution of

$$(2.39) \quad \frac{\partial^2 w}{\partial t^2} - \Delta w = 0 \quad \text{in } R^n \times [0, \infty),$$

$$(2.40) \quad w(z, 0) = Eu_0, \quad \frac{\partial w}{\partial t}(z, 0) = Ev_0.$$

We next choose $\psi \in C_0^\infty(\mathcal{O}_{2\delta})$ such that

$$(2.41) \quad \psi(z) = 1 \quad \text{for } z \in \mathcal{O}_\delta.$$

By solving the backward and forward Cauchy problem, we can obtain a solution $\Xi(z, t)$ of (2.39) in $R^n \times [0, \infty)$ which satisfies

$$(2.42) \quad \Xi(z, T) = \psi(z)w(z, T), \quad \frac{\partial \Xi}{\partial t}(z, T) = \psi(z) \frac{\partial w}{\partial t}(z, T).$$

Since $T \geq T_0$, we have

$$(2.43) \quad \Xi(z, T) \in C_0^\infty(\mathcal{O}_{2\delta}), \quad \frac{\partial \Xi}{\partial t}(z, T) \in C_0^\infty(\mathcal{O}_{2\delta})$$

from which it follows that

$$(2.44) \quad \Xi \in C^\infty(R^n \times [0, \infty)) \cap C([0, \infty); H^s(R^n)) \cap C^1([0, \infty); H^{s-1}(R^n)).$$

Let (\hat{u}_0, \hat{v}_0) be the restriction of $(\Xi(z, 0), \partial\Xi/\partial t(z, 0))$ to \mathcal{O}_δ . Then, it is evident that

$$(2.45) \quad (\hat{u}_0, \hat{v}_0) \in C^\infty(\overline{\mathcal{O}_\delta}) \times C^\infty(\overline{\mathcal{O}_\delta}).$$

Let us define the operator Λ_T , for $T \geq T_0$, by

$$(2.46) \quad \Lambda_T(u_0, v_0) = (\hat{u}_0, \hat{v}_0).$$

Then, Λ_T is a bounded linear operator from $H^s(\mathcal{O}_\delta) \times H^{s-1}(\mathcal{O}_\delta)$ into itself.

Lagnese [8] showed that for all $T \geq T_0$, possibly except for a finite number of values, $(I - \Lambda_T)^{-1}$ exists. Therefore, for given $\tilde{T} > T_0$, we can find T such that

$$(2.47) \quad \max \left\{ T_0, \tilde{T} - \frac{1}{2} \text{distance}(\partial\mathcal{O}, \partial\mathcal{O}_\delta) \right\} < T < \tilde{T}$$

and

$$(2.48) \quad (I - \Lambda_T)^{-1} \text{ exists.}$$

Hence, for given $(u_0^*, v_0^*) \in H^s(\mathcal{O}_\delta) \times H^{s-1}(\mathcal{O}_\delta)$, there is $(u_0, v_0) \in H^s(\mathcal{O}_\delta) \times H^{s-1}(\mathcal{O}_\delta)$ such that

$$(2.49) \quad (u_0, v_0) - \Lambda_T(u_0, v_0) = (u_0^*, v_0^*).$$

We then set

$$(2.50) \quad u = w - \Xi$$

where w is the solution of (2.39) and (2.40), and Ξ is the solution of (2.39) and (2.42). Then, it follows that

$$(2.51) \quad u \in C([0, \infty); H^s(R^n)) \cap C^1([0, \infty); H^{s-1}(R^n)),$$

$$(2.52) \quad u(z, 0) = u_0^*(z), \quad \frac{\partial u}{\partial t}(z, 0) = v_0^*(z) \quad \text{for } z \in \mathcal{O}_\delta,$$

$$(2.53) \quad u(z, T) = 0, \quad \frac{\partial u}{\partial t}(z, T) = 0 \quad \text{for } z \in \mathcal{O}_\delta,$$

which combined with (2.47) yields

$$(2.54) \quad u(z, \tilde{T}) = 0, \quad \frac{\partial u}{\partial t}(z, \tilde{T}) = 0 \quad \text{for } z \in \mathcal{O},$$

since the wave speed is 1. Now we assume that \mathcal{O} is convex and that $\tilde{T} > \text{diameter of } \mathcal{O}$ is given. Then, there are positive numbers δ, T , and T_0 such that (2.32), (2.33), (2.34), (2.38), (2.47), and (2.48) hold. We also assume that $\Theta = \text{supp } u_0^* \cup \text{supp } v_0^*$ is a compact subset of \mathcal{O} . Recalling that E is a total extension operator for \mathcal{O}_δ , we use (2.45) and (2.49) to find that

$$(2.55) \quad (Eu_0, Ev_0) \in C^\infty(R^n \setminus \Theta) \times C^\infty(R^n \setminus \Theta).$$

We can now apply Theorem 1.1 to the trace regularity of w on $\partial\mathcal{O} \times (0, \tilde{T})$. The procedure is the same as in §2.1 and we will omit the details. The control is given by

$$(2.56) \quad F(z, t) = \alpha u(z, t) + \beta \frac{\partial u}{\partial \nu}(z, t) \quad \text{on } \partial\mathcal{O} \times (0, \tilde{T}),$$

where α, β are nonnegative constants such that $\alpha + \beta > 0$.

THEOREM 2.7. *If $s \geq 2$, \mathcal{O} is a convex bounded domain with smooth boundary and $(u_0^*, v_0^*) \in H^s(\mathcal{O}) \times H^{s-1}(\mathcal{O})$ has compact support in \mathcal{O} , then the control given by (2.56) has the following property:*

$$(2.57) \quad F \in H^{s-1}(\partial\mathcal{O} \times (0, \tilde{T})) \quad \text{if } \beta \neq 0,$$

$$(2.58) \quad F \in H^s(\partial\mathcal{O} \times (0, \tilde{T})) \quad \text{if } \beta = 0.$$

Next we will consider the case $s < 2$. This case was not mentioned in [8], but we can easily infer the following argument.

Let $(u_0, v_0) \in H^s(\mathcal{O}) \times H^{s-1}(\mathcal{O})$ be given with compact support in \mathcal{O} . We then extend (u_0, v_0) so that $(u_0, v_0) \in H^s(R^n) \times H^{s-1}(R^n)$ by setting $u_0(z) = 0, v_0(z) = 0$ for $z \notin \mathcal{O}$. Then, we find a solution w of (2.39) in $R^n \times [0, \infty)$ with initial data (u_0, v_0) at $t = 0$. Suppose that $T > \text{diameter of } \mathcal{O}$. Then, it follows that

$$(2.59) \quad w(z, T) \in C^\infty(\overline{\mathcal{O}}), \quad \frac{\partial w}{\partial t}(z, T) \in C^\infty(\overline{\mathcal{O}}).$$

Since the controllability for the case $s \geq 2$ has been already established in [8] without any assumption on the support of the initial data, we may assert that there is $h \in H^2(\partial\mathcal{O} \times (0, T))$ such that the solution of

$$(2.60) \quad \frac{\partial^2 \Xi}{\partial t^2} - \Delta \Xi = 0 \quad \text{in } \mathcal{O} \times (0, T),$$

$$(2.61) \quad \Xi(z, T) = w(z, T), \quad \frac{\partial \Xi}{\partial t}(z, T) = \frac{\partial w}{\partial t}(z, T) \quad \text{in } \mathcal{O},$$

$$(2.62) \quad \alpha \Xi + \beta \frac{\partial \Xi}{\partial \nu} = h \quad \text{on } \partial\mathcal{O} \times (0, T)$$

satisfies

$$(2.63) \quad \Xi(z, 0) = 0, \quad \frac{\partial \Xi}{\partial t}(z, 0) = 0 \quad \text{in } \mathcal{O}.$$

We set

$$(2.64) \quad u = w - \Xi$$

so that u satisfies

$$(2.65) \quad \frac{\partial^2 u}{\partial t^2} - \Delta u = 0 \quad \text{in } \mathcal{O} \times (0, T),$$

$$(2.66) \quad u(z, T) = 0, \quad \frac{\partial u}{\partial t}(z, T) = 0 \quad \text{in } \mathcal{O}$$

$$(2.67) \quad u(z, 0) = u_0(z), \quad \frac{\partial u}{\partial t}(z, 0) = v_0(z) \quad \text{in } \mathcal{O}.$$

Hence, a desired control F is given by

$$(2.68) \quad F = \alpha u + \beta \frac{\partial u}{\partial \nu} \quad \text{on } \partial \mathcal{O} \times (0, T).$$

For the regularity of F , we only need to consider the trace regularity of w since h in (2.62) is smooth enough. By the same argument as in the previous section, we can assert the following results under the assumptions:

- (i) \mathcal{O} is a convex bounded domain with smooth boundary;
 - (ii) $T > \text{diameter of } \mathcal{O}$;
 - (iii) $(u_0, v_0) \in H^s(\mathcal{O}) \times H^{s-1}(\mathcal{O})$ with compact support in \mathcal{O} .
- THEOREM 2.8. *If $1 \leq s < 2$, then the control F given by (2.68) satisfies*

$$(2.69) \quad F \in H^{s-1}(\partial \mathcal{O} \times (0, T)) \quad \text{if } \beta \neq 0,$$

$$(2.70) \quad F \in H^s(\partial \mathcal{O} \times (0, T)) \quad \text{if } \beta = 0.$$

If $s < 1$ and $\beta = 0$, then (2.70) holds, and if $s < 1$ and $\alpha = 0$, then (2.69) holds.

For $s < 1$, we used the following definition of weak solution, which is a slight variant of Definition 2.3.

DEFINITION 2.9. Suppose that K is a compact subset of \mathcal{O} and that ρ is a positive number such that $\rho < \frac{1}{2} \text{ distance } (\partial \mathcal{O}, K)$. Let $s \in \mathbb{R}$, $(u_0, v_0) \in H_K^s(\mathbb{R}^n) \times H_K^{s-1}(\mathbb{R}^n)$ and $g = g_1 + g_2$, where $g_1 \in \tilde{H}_\rho^s(\partial \mathcal{O} \times (0, T))$ and $g_2 \in L^2(\partial \mathcal{O} \times (0, T))$. Then, a function $u(z, t) \in C([0, T]; H^s(\mathcal{O})) \cap C^1([0, T]; H^{s-1}(\mathcal{O}))$ is called a solution of the initial-boundary value problem (2.1), (2.2), and (2.22) if u satisfies

$$(2.71) \quad \begin{aligned} \int_0^T \langle u(z, t), \chi(z, t) \rangle_1 dt &= \langle v_0(z), \theta(z, 0) \rangle_2 - \left\langle u_0(z), \frac{\partial \theta}{\partial t}(z, 0) \right\rangle_3 - \left\langle g_1(z, t), \frac{\partial \theta}{\partial \nu}(z, t) \right\rangle_4 \\ &\quad - \int_0^T \int_{\partial \mathcal{O}} g_2(z, t) \frac{\partial \theta}{\partial \nu}(z, t) dz dt, \end{aligned}$$

for each $\chi \in C_0^\infty(\mathcal{O} \times (0, T))$, where $\theta(z, t) \in C^\infty(\overline{\mathcal{O}} \times [0, T])$ is the solution of (2.24), (2.25), and (2.26), and we have used the notation (2.27)–(2.30).

In the above definition, it is necessary to split g into two parts because Ξ , which is the smoother part of u in (2.64), does not vanish near $t = 0$ and does not belong to $\tilde{H}_\rho^s(\partial \mathcal{O} \times (0, T))$. Next we observe that if $g = g_1 + g_2 = \tilde{g}_1 + \tilde{g}_2$ such that $g_1, \tilde{g}_1 \in \tilde{H}_\rho^s(\partial \mathcal{O} \times (0, T))$ and $g_2, \tilde{g}_2 \in L^2(\partial \mathcal{O} \times (0, T))$, then

$$(2.72) \quad \begin{aligned} \langle g_1 - \tilde{g}_1, \theta \rangle_4 &= \int_0^T \int_{\partial \mathcal{O}} (g_1 - \tilde{g}_1) \theta dz dt \\ &= \int_0^T \int_{\partial \mathcal{O}} (\tilde{g}_2 - g_2) \theta dz dt \end{aligned}$$

holds for each $\theta \in C^\infty(\partial \mathcal{O} \times (0, T))$ which vanishes near $t = T$, since $g_1 - \tilde{g}_1 \in L^2(\partial \mathcal{O} \times (0, T))$ and $\langle \cdot, \cdot \rangle_4$ reduces to the L^2 inner product.

Hence,

$$(2.73) \quad \langle g_1, \theta \rangle_4 + \int_0^T \int_{\partial \mathcal{O}} g_2 \theta dz dt = \langle \tilde{g}_1, \theta \rangle_4 + \int_0^T \int_{\partial \mathcal{O}} \tilde{g}_2 \theta dz dt$$

holds and (2.71) is independent of the decomposition of g .

3. Application to the Neumann boundary value problem. In this section, we will discuss the following initial-boundary value problem for a second-order hyperbolic equation.

$$(3.1) \quad \begin{aligned} c_0(x, y, t) \frac{\partial^2 u}{\partial t^2} &= \sum_{i,j=1}^{n-1} a_{ij}(x, y, t) \frac{\partial^2 u}{\partial x_i \partial x_j} + \frac{\partial^2 u}{\partial y^2} \\ &+ \sum_{i=1}^{n-1} b_i(x, y, t) \frac{\partial u}{\partial x_i} + b_0(x, y, t)u + c_1(x, y, t) \frac{\partial u}{\partial t} \\ &+ d(x, y, t) \frac{\partial u}{\partial y} + f(x, y, t), \quad \text{for } (x, y, t) \in R^{n-1} \times R^+ \times R^+, \end{aligned}$$

$$(3.2) \quad \frac{\partial u}{\partial y}(x, 0, t) = 0 \quad \text{for } (x, t) \in R^{n-1} \times R^+$$

$$(3.3) \quad u(x, y, 0) = u_0(x, y), \quad \frac{\partial u}{\partial t}(x, y, 0) = v_0(x, y) \quad \text{for } (x, y) \in R^{n-1} \times R^+,$$

where $R^+ = (0, \infty)$ and $n \geq 2$.

We assume that

$$(3.4) \quad a'_{ij}s, b'_is, c'_is \text{ and } d \text{ belong to } C^\infty(R^{n+1}) \text{ and these functions together with all their derivatives are bounded in } R^{n+1},$$

$$(3.5) \quad \begin{aligned} a_{ij} &= a_{ji} \quad \text{and} \quad \sum_{i,j=1}^{n-1} a_{ij}(x, y, t) \xi_i \xi_j \geq k_1 |\xi|^2 \quad \text{for all } \xi \in R^{n-1} \\ &\text{and all } (x, y, t) \in R^{n-1} \times R \times R, \end{aligned}$$

where k_1 is a positive constant,

$$(3.6) \quad c_0(x, y, t) \geq k_2 \quad \text{for all } (x, y, t) \in R^{n-1} \times R \times R,$$

where k_2 is positive constant,

$$(3.7) \quad \begin{aligned} a'_{ij}s, b'_is \text{ and } c'_is &\text{ are even functions with respect to the } y \text{ variable and } d \\ &\text{ is an odd function with respect to the } y \text{ variable.} \end{aligned}$$

THEOREM 3.1. Suppose that $u_0(x, y) \in H^1(R^{n-1} \times R^+)$ and $v_0(x, y) \in L^2(R^{n-1} \times R^+)$ have compact support in $R^{n-1} \times R^+$ and $f(x, y, t) \in L^2(R^{n-1} \times R^+ \times R^+)$ has compact support in $R^{n-1} \times R^+ \times \overline{R^+}$. Let $u \in C([0, \infty); H^1(R^{n-1} \times R^+)) \cap C^1([0, \infty); L^2(R^{n-1} \times R^+))$ be the unique solution of (3.1), (3.2), and (3.3). Then, for each $T > 0$, we have

$$(3.8) \quad u(x, 0, t) \in H^1(R^{n-1} \times (0, T)).$$

Proof. Since u_0 , v_0 , and f vanish near $y = 0$, we can easily extend u_0 , v_0 , and f so that

$$(3.9) \quad u_0, v_0, \text{ and } f \text{ are even functions with respect to the } y \text{ variable,}$$

$$(3.10) \quad f(x, y, t) = 0 \quad \text{for all } (x, y) \in R^{n-1} \times R \quad \text{if } t < 0.$$

It follows that

$$(3.11) \quad (u_0, v_0) \in H^1(R^n) \times L^2(R^n),$$

$$(3.12) \quad f \in L^2(R^{n+1}).$$

Let $u \in C(R; H^1(R^n)) \cap C^1(R; L^2(R^n))$ be the solution of the pure Cauchy problem for (3.1) in $R^n \times R$ satisfying (3.3) in R^n . By virtue of (3.7) and (3.9), u is an even function with respect to the y variable and consequently,

$$(3.13) \quad \frac{\partial u}{\partial y}(x, 0, t) = 0 \quad \text{for almost all } (x, t) \in R^{n-1} \times R.$$

By the uniqueness of solution to the initial-boundary value problem (3.1), (3.2), and (3.3), the restriction of u to $R^{n-1} \times R^+ \times R^+$ coincides with the solution of (3.1), (3.2), and (3.3). Furthermore, because of finite wave speed, it is enough to establish local trace regularity. Hence, the question of trace regularity is posed in the same format as in Theorem 1.1 with $s = 1$ and $y_0 = 0$.

It remains to verify the condition (1.4). The principal symbol of (3.1) is given by

$$(3.14) \quad P(x, y, t, \xi, \eta, \tau) = \eta^2 + \sum_{i,j=1}^{n-1} a_{ij}(x, y, t) \xi_i \xi_j - c_0(x, y, t) \tau^2$$

where $(\xi, \eta, \tau) \in R^{n-1} \times R \times R$ denotes the dual variables corresponding to $(x, y, t) \in R^{n-1} \times R \times R$. Let $(x(\alpha), y(\alpha), t(\alpha), \xi(\alpha), \eta(\alpha), \tau(\alpha))$ for $\alpha \in R$ be a bicharacteristic strip of (3.14) such that

$$(3.15) \quad \begin{aligned} x(0) &= x_0, & y(0) &= 0, & t(0) &= t_0 \\ \xi(0) &= \xi_0 \neq 0, & \eta(0) &= 0, & \tau(0) &= \tau_0 \neq 0. \end{aligned}$$

Then,

$$(3.16) \quad P(x_0, 0, t_0, \xi_0, 0, \tau_0) = 0$$

and

$$(3.17) \quad \frac{dx_i}{d\alpha} = \frac{\partial}{\partial \xi_i} P(x, y, t, \xi, \eta, \tau), \quad i = 1, \dots, n-1,$$

$$(3.18) \quad \frac{dy}{d\alpha} = \frac{\partial}{\partial \eta} P(x, y, t, \xi, \eta, \tau),$$

$$(3.19) \quad \frac{dt}{d\alpha} = \frac{\partial}{\partial \tau} P(x, y, t, \xi, \eta, \tau),$$

$$(3.20) \quad \frac{d\xi_i}{d\alpha} = -\frac{\partial}{\partial x_i} P(x, y, t, \xi, \eta, \tau), \quad i = 1, \dots, n-1$$

$$(3.21) \quad \frac{d\eta}{d\alpha} = -\frac{\partial}{\partial y} P(x, y, t, \xi, \eta, \tau),$$

$$(3.22) \quad \frac{d\tau}{d\alpha} = -\frac{\partial}{\partial t} P(x, y, t, \xi, \eta, \tau).$$

Since a'_{ij} s and c_0 are even functions with respect to the y variable, we have

$$(3.23) \quad \frac{\partial}{\partial y} P(x, 0, t, \xi, \eta, \tau) = 0 \quad \text{for every } x, t, \xi, \eta \text{ and } \tau.$$

Consequently, it follows from (3.15), (3.18), (3.21), and (3.23) that

$$(3.24) \quad y(\alpha) = 0, \quad \eta(\alpha) = 0 \quad \text{for all } \alpha$$

of the interval where the bicharacteristic strip is defined.

Next, we will observe the following fact.

LEMMA 3.2. *Let $t_0 \geq 0$. If $\tau_0 = \pm 1$, then there are positive constants α_0 and ϵ_0 independent of x_0, t_0 , and ξ_0 satisfying (3.16) such that the bicharacteristic strip is defined on $[-\alpha_0, \alpha_0]$ and such that $t(\alpha_0) \leq t_0 - \epsilon_0$ for $\tau_0 = 1$ and $t(-\alpha_0) \leq t_0 - \epsilon_0$ for $\tau_0 = -1$. Furthermore, if $\tau_0 = 1$, then $\tau(\alpha) > 0$ for all $\alpha \in [-\alpha_0, \alpha_0]$ and if $\tau_0 = -1$, then $\tau(\alpha) < 0$ for all $\alpha \in [-\alpha_0, \alpha_0]$.*

Proof. We first note that

$$(3.25) \quad P(x(\alpha), 0, t(\alpha), \xi(\alpha), 0, \tau(\alpha)) = 0$$

holds along the bicharacteristic strip, and rewrite (3.17), (3.19), (3.20), and (3.22) as

$$(3.26) \quad \frac{dx_i}{d\alpha} = 2 \sum_{j=1}^{n-1} a_{ij}(x, 0, t) \xi_j, \quad i = 1, \dots, n-1,$$

$$(3.27) \quad \frac{dt}{d\alpha} = -2c_0(x, 0, t)\tau,$$

$$(3.28) \quad \frac{d\xi_i}{d\alpha} = - \sum_{j,k=1}^{n-1} \frac{\partial}{\partial x_i} a_{jk}(x, 0, t) \xi_j \xi_k + \frac{\partial}{\partial x_i} c_0(x, 0, t) \tau^2, \quad i = 1, \dots, n-1,$$

$$(3.29) \quad \frac{d\tau}{d\alpha} = - \sum_{j,k=1}^{n-1} \frac{\partial}{\partial t} a_{jk}(x, 0, t) \xi_j \xi_k + \frac{\partial}{\partial t} c_0(x, 0, t) \tau^2.$$

Since ξ_0 satisfies (3.16), we infer from (3.4) and (3.5) that for $\tau_0 = \pm 1$,

$$(3.30) \quad |\xi_0| \leq M_1,$$

holds for some positive constant M_1 which is independent of x_0 and t_0 . Hence, by (3.4) and (3.30), there are positive constants α_1 and M_2 independent of x_0, t_0 , and ξ_0 such that the bicharacteristic strip is defined on $[-\alpha_1, \alpha_1]$ and

$$(3.31) \quad |\tau(\alpha)| + |\xi(\alpha)| \leq M_2$$

for all $\alpha \in [-\alpha_1, \alpha_1]$.

It follows from (3.4), (3.29), and (3.31) that

$$(3.32) \quad \left| \frac{d\tau}{d\alpha} \right| \leq M_3 \quad \text{for all } \alpha \in [-\alpha_1, \alpha_1]$$

for some positive constant M_3 independent of x_0, t_0 , and ξ_0 . Now we use (3.6), (3.27), and (3.32) to find $0 < \alpha_0 \leq \alpha_1$ and $\epsilon_0 > 0$ as in the assertion of the lemma. For the last statement of the lemma, we argue as follows. If $\tau(\alpha_2) = 0$, for some $\alpha_2 \in [-\alpha_1, \alpha_1]$, then

$\xi(\alpha_2) = 0$ by (3.25). As a result of the uniqueness of solution to the system (3.26)–(3.29), it is evident that $x(\alpha) = x(\alpha_2)$, $t(\alpha) = t(\alpha_2)$, $\tau(\alpha) = 0$, and $\xi(\alpha) = 0$ should hold for all $\alpha \in [-\alpha_1, \alpha_1]$. This contradicts $\tau(0) = \pm 1$. Now the proof of the lemma is complete.

We also need the following fact.

LEMMA 3.3. *There is a constant $\epsilon_1 > 0$ such that*

$$(3.33) \quad \{(x, y, t) : x \in R^{n-1}, y = 0, -\epsilon_1 < t < 0\} \\ \text{is disjoint from the singular support of } u.$$

Proof. We recall (3.10) and the assumption that

$$(3.34) \quad \{(x, y) : x \in R^{n-1}, y = 0\} \text{ is disjoint from } \text{supp } u_0 \cup \text{supp } v_0.$$

The assertion (3.33) is an immediate consequence of finite domain of dependence.

We now verify condition (1.4). Choose any $(x_0, t_0) \in R^{n-1} \times R^+$ and $\xi_0 \in R^{n-1}$ which satisfies (3.16) with $\tau_0 = 1$. Since $\{(x, y, t) : x \in R^{n-1}, y = 0, t \in R\}$ is disjoint from $\text{supp } f$, Proposition 3.5.1 of [5] implies that $(x_0, 0, t_0, \xi_0, 0, 1) \notin WF(u)$ if and only if $(x(\alpha), 0, t(\alpha), \xi(\alpha), 0, \tau(\alpha)) \notin WF(u)$ for some $\alpha \in [-\alpha_0, \alpha_0]$. Here α_0 is the same as in Lemma 3.2, which says that $t(\alpha_0) \leq t_0 - \epsilon_0$. Thus, if $t(\alpha_0) < 0$, then $(x_0, 0, t_0, \xi_0, 0, 1) \notin WF(u)$ according to Lemma 3.3. If $0 \leq t(\alpha_0) \leq t_0 - \epsilon_0$, we argue as follows. Since $\tau(\alpha_0) > 0$, it is evident that $(x(\alpha_0), 0, t(\alpha_0), \xi(\alpha_0), 0, \tau(\alpha_0)) \notin WF(u)$ if and only if $(x(\alpha_0), 0, t(\alpha_0), \xi(\alpha_0)/\tau(\alpha_0), 0, 1) \notin WF(u)$. By taking $(x(\alpha_0), 0, t(\alpha_0), \xi(\alpha_0)/\tau(\alpha_0), 0, 1)$ as the initial value of the bicharacteristic strip, we apply Lemma 3.2 so that the new terminal value of t satisfies

$$(3.35) \quad t \leq t_0 - 2\epsilon_0.$$

If $t_0 - 2\epsilon_0 < 0$, then obviously, $(x_0, 0, t_0, \xi_0, 0, 1) \notin WF(u)$. If not, we repeat the same procedure until we reach negative value of t . For $\tau_0 = -1$, we use the same argument to find $(x_0, 0, t_0, \xi_0, -1) \notin WF(u)$. Now condition (1.4) has been established, which completes the proof of Theorem 3.1.

Remark 3.4. When (3.1) reduces to a wave equation with constant coefficients, the above result was already obtained by Symes [18] and Lasiecka and Triggiani [10]. The method of geometric optics was used in [18] and the Fourier–Laplace transform was used in [10]. In [10], we can find a very enlightening example which shows that the H^1 regularity is not true, in general, when the support of f intersects the boundary. For the general second-order hyperbolic operator, the most up-to-date results on the trace regularity of solution to the Neumann boundary value problem can be found in [11] and [12], where it is not assumed that the data have compact support.

REFERENCES

- [1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] G. BAO AND W. W. SYMES, *A trace theorem for solutions of linear partial differential equations*, Math. Methods Appl. Sci., 14 (1991), pp. 553–562.
- [3] ———, *Trace regularity for a second order hyperbolic equation with nonsmooth coefficients*, preprint.
- [4] C. BARDOS, G. LEBEAU, AND J. RAUCH, *Contrôle et stabilisation dans les problèmes hyperboliques*, Appendix II, in Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués, J. L. Lions, ed., Masson, Paris, 1988.
- [5] L. HÖRMANDER, *On the existence and the regularity of solutions of linear pseudodifferential equations*, L'Enseignement Math., 17 (1971), pp. 99–163.
- [6] ———, *The Analysis of Linear Partial Differential Operators III*, Springer-Verlag, Berlin, New York, 1985.

- [7] M. A. HORN AND I. LASIECKA, *Asymptotic behavior with respect to thickness of boundary stabilizing feedback for the Kirchhoff plate*, preprint.
- [8] J. LAGNESE, *Boundary value control of a class of hyperbolic equations in a general region*, SIAM J. Control Optim. 15 (1977), pp. 973–983.
- [9] ———, *On the support of solutions of the wave equation with applications to exact boundary value controllability*, J. Math. Pures et. Appl., 58 (1979), pp. 121–135.
- [10] I. LASIECKA AND R. TRIGGIANI, *Trace regularity of the solutions of the wave equation with homogeneous Neumann boundary conditions and compactly supported data*, J. Math. Anal. Appl., 141 (1989), pp. 49–71.
- [11] I. LASIECKA AND R. TRIGGIANI, *Sharp regularity theory for second order hyperbolic equations of Neumann type*, Ann. Mat. Pura Appl., CLVII (1990), pp. 285–367.
- [12] ———, *Regularity theory of hyperbolic equations with nonhomogeneous Neumann boundary conditions, Part II*, J. Differential Equations, 94 (1991), pp. 112–164.
- [13] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, vol. 1, Springer-Verlag, Berlin, New York, 1972.
- [14] W. LITTMAN, *Boundary control theory for partial differential equations*, Ann. Scuola Norm. Sup. Pisa, 5 (1978), pp. 567–580.
- [15] ———, *Near optimal time boundary controllability for a class of hyperbolic equations*, in Lecture Notes in Control and Information Science, No. 97, Springer-Verlag, Berlin, New York, 1987, pp. 307–312.
- [16] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.
- [17] ———, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev. 20 (1978), pp. 639–739.
- [18] W. W. SYMES, *A trace theorem for solutions of the wave equation and the remote determination of acoustic sources*, Math. Methods Appl. Sci., 5 (1983), pp. 35–93.
- [19] M. TAYLOR, *Pseudodifferential Operators*, Princeton University Press, Princeton, NJ, 1981.

STATE SPACE REALIZATION OF 2-D FINITE-DIMENSIONAL BEHAVIOURS*

E. FORNASINI[†], P. ROCHA[‡], AND S. ZAMPIERI[†]

Abstract. This paper deals with the state space realization of autonomous autoregressive two-dimensional (2-D) systems in the context of the behavioural approach. An arbitrary autoregressive 2-D system Σ can be viewed as the sum of an externally controllable subsystem Σ^c with an autonomous one Σ^a , so that a state space realization of Σ can be obtained by separately realizing Σ^c and Σ^a . Since a procedure for realizing externally controllable systems in state/driving-variable form is already available in the literature, the general realization problem is easily reduced to the autonomous case. Here, some properties of finite-dimensional autonomous systems are discussed, allowing for a realization procedure that uses the Gröbner bases theory.

Key words. autonomous systems, (externally) controllable systems, Gröbner basis, state/driving-variable realization

AMS subject classifications. 93B20, 93B25

1. Introduction. In this paper we will present some results and algorithms involved in the construction of two-dimensional (2-D) state space models on the basis of external data. Following the behavioural approach to dynamical systems introduced in [1], [2], [3], external data are characterized by means of a family of laws telling us that certain signals can occur and others cannot. Moreover, all the components of the external data play completely symmetric roles, so that no input/output structure is a priori assumed.

Realization theory has been developed for the most part in the one-dimensional environment, where state space models have shown to be a very convenient framework for the mathematical analysis and synthesis of real time data processors and controllers. In this context the state is very naturally viewed as a set of latent variables which parametrize the content of system memory, and the realization problems are inextricably connected with the definitions of past and future that underlie the notion of memory.

When trying to formulate state concepts for 2-D systems, it is of central importance to realize that there is often no natural, intrinsic direction of the evolution for systems defined over a 2-D domain. In this case any choice of a preferred direction is artificial, and there are various possible definitions of past and future. Some of them are discussed in [1], where their connections with 2-D state representations available in the literature are illustrated in detail. Obviously, the most classical example is provided by the quarter plane causality structure. It underlies the so-called South West (SW)-state representation, which is the behavioural counterpart of the state space model of quarter plane impulse responses, introduced by Attasi [4], Roesser [5], and Fornasini and Marchesini [6] within the classical input/output framework.

Singular state space models have been analyzed to cope with more general causality structures in, e.g., Morf [7], Kaczorek [8], and Lewis [9]. In this paper an alternative approach to noncausal structures is considered, based on the introduction of a set of auxiliary free variables in SW-state representation. These act in some sense like an input driving the system dynamics and are therefore called the “driving variables.” State/driving-variable models allow us to compute recursively joint input-output trajectories from the values of the auxiliary free variables via the state. Although only the realization of 2-D “auto-regressive (AR)” equations is considered, we remark, however, that transfer functions can also be handled in this context, since they can be identified with (externally) controllable AR systems; see [1].

It is shown in [1] that every controllable AR 2-D system (and hence every 2-D rational transfer function) can be realized in SW-state/driving-variable form, independently of the

* Received by the editors December 30, 1991; accepted for publication (in revised form) August 12, 1992.

[†] Dipartimento di Elettronica ed Informatica, Università di Padova, via Gradenigo 6/a, 35131 Padova, Italia.

[‡] Faculty of Technical Mathematics and Informatics, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, the Netherlands.

existence of a quarter plane causal structure between the system variables. This results from the possibility of generating the joint input-output trajectories $w = \text{col}(u, y)$ of a controllable AR system from (auxiliary) driving-variable trajectories v such that the relationship between w and v is quarter plane causal, even if u and y are not causally related. Viewing v as the new “input” and w as the new “output,” a 2-D state space model (for instance, of the Fornasini–Marchesini (FM) type [6]) can then be obtained by the classical realization procedures, yielding the desired state/driving-variable realization.

The realizability of arbitrary, i.e., not necessarily controllable, AR systems is still an open problem. Reducing this problem to the realization of autonomous systems is the first goal of our paper. Indeed, we show that every AR system Σ can be viewed as the sum of an externally controllable system Σ^c and an autonomous one Σ^a and, therefore, the realization of Σ can be obtained by realizing separately Σ^c and Σ^a . Since a procedure for realizing externally controllable systems is given in [1], it will be enough to derive a realization procedure for autonomous systems.

This contribution focuses on finite-dimensional autonomous AR systems with q (real-valued) variables defined over \mathbb{Z}^2 . It turns out that, in this case, the admissible system trajectories constitute a finite-dimensional vector space \mathcal{B} , given by $\mathcal{B} = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q | R(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0\}$, with σ_1 and σ_2 the 2-D shifts and $R(z_1, z_2, z_1^{-1}, z_2^{-1})$ a factor right prime 2-D polynomial matrix.

The second goal we pursue in this paper is that of representing the autonomous behaviour via a state model, characterized by a pair of commuting matrices that describe the state evolution in the two directions of the grid. The realization algorithm exploits the algebraic duality between \mathcal{B} and a suitable quotient module over the space of 2-D polynomial rows, as well as the correspondence between the shift operators in \mathcal{B} and a pair of adjoint operators in the quotient module. These operators are represented by a pair of commutative invertible matrices and can be obtained by computer algebra techniques and linear manipulations.

2. Autonomous 2-D systems. Following the behavioural approach to dynamical systems introduced in [1] and [2], we characterize a 2-D system by means of its behaviour, which consists of the set of all the signals which are compatible with the system laws. Moreover, we do not start with a given input/output structure, i.e., the system signals are stacked together in a signal w instead of being divided into inputs u and outputs y . A 2-D system Σ with q real valued variables defined over \mathbb{Z}^2 and with behaviour $\mathcal{B} \subseteq \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q\}$ will be denoted by $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$.

In the sequel we will be interested in the class of autoregressive 2-D systems. $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ is said to be an AR system if there exists a 2-D Laurent polynomial matrix $R(z_1, z_2, z_1^{-1}, z_2^{-1})$ such that

$$\mathcal{B} = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q | R(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0\} =: \ker R(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1}),$$

with σ_1 and σ_2 the 2-D shift operators. An AR 2-D system is said to be finite-dimensional if its behaviour \mathcal{B} is a finite-dimensional vector space; otherwise, Σ is said to be infinite-dimensional.

In order to define the notion of autonomy we introduce the following nomenclature. A subset of \mathbb{R}^2 is said to be 2-D unbounded if it contains a plane sector $\mathcal{S}(v, v_1, v_2) := \{v + \alpha v_1 + \beta v_2 | \alpha, \beta \geq 0\}$ with $v, v_1, v_2 \in \mathbb{R}^2$, and v_1, v_2 linearly independent. $U \subseteq \mathbb{Z}^2$ is a 2-D unbounded set if $U = \mathcal{U} \cap \mathbb{Z}^2$ for some 2-D unbounded set \mathcal{U} in \mathbb{R}^2 .

Definition 1. $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ is an autonomous 2-D system if there exists a subset $T \subseteq \mathbb{Z}^2$ such that $\mathbb{Z}^2 \setminus T$ is 2-D unbounded and satisfies the condition $\{w_1, w_2 \in \mathcal{B} \text{ and } w_1|_T = w_2|_T\} \Rightarrow \{w_1 = w_2\}$.

So, intuitively, a system is autonomous if the evolution of its trajectories in a sufficiently large portion $\mathbb{Z}^2 \setminus T$ of the discrete plane is completely specified by what occurs in the remaining portion T of the domain \mathbb{Z}^2 . As stated in Proposition 2.1, for autoregressive systems the autonomy is equivalent to the absence of free variables.

Notation. Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ be a system in the variables $(w_1, \dots, w_q)^T := w$. The variable $w_i, i \in \{1, \dots, q\}$, is a free variable if, for every $\alpha : \mathbb{Z}^2 \rightarrow \mathbb{R}$, there exists some $w \in \mathcal{B}$ such that $w_i = \alpha$. Similarly, a vector $(w_{i_1}, \dots, w_{i_l})^T$, with $i_j \in J \subseteq \{1, \dots, q\}$ and $i_j \neq i_k$ if $j \neq k$, is a vector of free variables if for every $\beta : \mathbb{Z}^2 \rightarrow \mathbb{R}^l$ there exists $w \in \mathcal{B}$ such that $(w_{i_1}, \dots, w_{i_l})^T = \beta$. The number of free variables in a system Σ is defined as the maximum dimension of a vector of free variables in Σ .

LEMMA 2.1. *Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ be an autoregressive 2-D system such that $\mathcal{B} = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q | R(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0\}$. Then*

1. *The number l of free variables in Σ is $q - \text{rank } R$;*

2. *If $l > 0$, there exists a nonzero $q \times l$ polynomial matrix $M(z_1, z_2, z_1^{-1}, z_2^{-1})$ such that $\mathcal{B} \supseteq \text{im } M(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})$, where $M(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})$ is viewed as an operator from $\{v : \mathbb{Z}^2 \rightarrow \mathbb{R}^l\}$ into $\{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q\}$. Moreover, $\text{im } M$ has l free variables.*

Proof. In order to prove statement 1, we first consider the case where R has full row rank. In this case, there is a column permutation Π such that $R\Pi = [P|Q]$ with P square $r \times r$ and nonsingular. This means that the equation $Rw = 0$ is equivalent to

$$(2.1) \quad Pw_1 = -Qw_2,$$

with $\text{col}(w_1, w_2) = \Pi w$. Since P is full row rank, it can be shown that $P(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})$ is a surjective operator, and w_2 is a $q - r$ -dimensional vector of free variables. So $l \geq q - r$. Now, it remains to see that none of the components of w_1 are free. Let $P^*(z_1, z_2, z_1^{-1}, z_2^{-1})$ be such that $P^*P = \text{diag}(p) =: D$, with $p := \det P$, and define $E := -P^*Q$. Premultiplying (2.1) by P^* yields

$$(2.2) \quad Dw_1 = Ew_2.$$

In particular, if $w_2 \equiv 0$, (2.2) implies that the components w_{1i} of w_1 must satisfy

$$(2.3) \quad p(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w_{1i} = 0 \quad i = 1, \dots, r$$

and are, hence, not free. This shows that the number l of free variables in $\mathcal{B} := \ker R$ is exactly $l = q - r$.

If R does not have full row rank, there exists a factorization

$$(2.4) \quad R = F\bar{R}$$

such that F has full column rank, \bar{R} has full row rank and $\text{rank } F = \text{rank } \bar{R} = \text{rank } R$. Let \bar{F} be an $r \times r$ submatrix of F obtained by selecting r linearly independent rows. It is not difficult to see that

$$\bar{R}w = 0 \Rightarrow F\bar{R}w = 0 \Rightarrow \bar{F}\bar{R}w = 0,$$

or, equivalently,

$$\mathcal{B}_1 := \ker \bar{R} \subseteq \mathcal{B} := \ker R \subseteq \ker \bar{F}\bar{R} =: \mathcal{B}_2.$$

Since \bar{R} and $\bar{F}\bar{R}$ are both matrices with full row rank r , it follows from the previous reasoning that both \mathcal{B}_1 and \mathcal{B}_2 have $l = q - r$ free variables, proving the first statement of the lemma.

As for statement 2, if $l = q - r$, without loss of generality, the matrix \bar{R} in the factorization (2.4) can be taken to be a full rank factor left prime 2-D polynomial matrix of size $r \times q$ (note that if \bar{R} is not left prime its nontrivial left factors can be extracted and included in F). In this case it follows from [1, Thm. 1] that there exists a $q \times l$ matrix $M(z_1, z_2, z_1^{-1}, z_2^{-1})$ such that $\ker \bar{R}(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1}) = \text{im } M(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})$. So, $\mathcal{B} \supseteq \ker \bar{R} = \text{im } M$. Finally, we note that the number of free variables in $\ker \bar{R}$ is still l , and hence $\text{im } M = \ker \bar{R}$ has l free variables. \square

PROPOSITION 2.1. *An autoregressive system $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ is autonomous if and only if it has no free variables.*

Proof. (i) Suppose that Σ is a system without free variables. This means that $\mathcal{B} = \ker R$, for some full rank matrix $R(z_1, z_2, z_1^{-1}, z_2^{-1})$. Let P be a $q \times q$ matrix obtained by taking q rows of R , and define $P^* := DP^{-1}$ with $D := \text{diag}(p)$ and $p := \det P$ (note that P^* is a polynomial matrix). Clearly,

$$Rw = 0 \Rightarrow Pw = 0 \Rightarrow Dw = 0 \Leftrightarrow p(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w_i = 0 \quad i = 1, \dots, q,$$

where w_i denotes the i th component of w .

We next show that $\ker D$ is an autonomous behaviour. Since $\mathcal{B} \subseteq \ker D$, this implies that also \mathcal{B} (and hence Σ) is autonomous. Since the support of p is finite, then there exist integers l_1, L_1, l_2, L_2 such that such support is included in the set $T := \{(h_1, h_2) \in \mathbb{Z}^2 \mid l_1 \leq h_1 \leq L_1 \text{ or } l_2 \leq h_2 \leq L_2\}$. It can be shown that the solutions of the equation are completely determined by their values on T , i.e., if w_1, w_2 are elements of $\ker D$ and $w_1|_T = w_2|_T$, then $w_1 = w_2$. Since $\mathbb{Z}^2 \setminus T$ is 2-D unbounded, this means that $\ker D$ is autonomous.

(ii) Assume that Σ has $l > 0$ free variables and let M be as in Lemma 2.1. Denote, respectively, by \bar{m}_i and \underline{m}_i the maximum and the minimum of the exponents of z_i in the entries of M , and define the extent of M as $e(M) := \sqrt{2} \max\{\bar{m}_1 - \underline{m}_1, \bar{m}_2 - \underline{m}_2\}$. Further, denote the Euclidean distance by $d(\cdot, \cdot)$. Given any subset $T \subseteq \mathbb{Z}^2$ such that $\mathbb{Z}^2 \setminus T$ is 2-D unbounded, define two trajectories v' and $v'' \in \{v : \mathbb{Z}^2 \rightarrow \mathbb{R}^q\}$ in the following way. The trajectory v' is simply the zero trajectory. As for v'' , we require that $v''(t_1, t_2) = 0$ if $d((t_1, t_2), T) \leq e(M)$; for (t_1, t_2) with $d((t_1, t_2), T) > e(M)$, we define $v''(t_1, t_2)$ in such a way that $Mv'' \neq 0$. Note that this is possible since the value of Mv'' at a point $(t_1^*, t_2^*) \in \mathbb{Z}^2$ depends only on the values of v'' at the points (t_1, t_2) such that $d((t_1, t_2), (t_1^*, t_2^*)) \leq e(M)$. Now let $w' := Mv' = 0$ and $w'' := Mv''$. Clearly, $w', w'' \in \mathcal{B}$. Moreover $w'|_T = w''|_T = 0$, and $w'' \neq 0 = w'$. This shows that Σ is not autonomous. \square

Contrary to the one-dimensional case, where autonomous linear systems are necessarily finite-dimensional [2], autonomous 2-D systems may be infinite-dimensional. The following proposition characterizes the autonomy and the finite dimensionality properties of AR 2-D systems.

PROPOSITION 2.2. *Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ be an AR 2-D system with behaviour $\mathcal{B} = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q \mid R(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0\}$, where $R(z_1, z_2, z_1^{-1}, z_2^{-1})$ is a 2-D polynomial matrix. Then Σ is autonomous if and only if R has full column rank. Moreover Σ is finite-dimensional if and only if R is right factor prime.*

Proof. The first part of the result is an immediate consequence of Lemma 2.1 and Proposition 2.1. To prove the second statement assume first that R is right factor prime. Without loss of generality, we can suppose that the entries of R are in $\mathbb{R}[z_1, z_2]$. Then there exist [7] matrices $X_i(z_1, z_2)$ such that

$$X_i(z_1, z_2)R(z_1, z_2) = \text{diag}(d(z_i), \dots, d(z_i)), \quad i = 1, 2.$$

So $Rw = 0$ implies $d(\sigma_i)w_j = 0, i = 1, 2$. Consequently the projection of \mathcal{B} into the j th signal component is finite-dimensional ($j = 1, \dots, q$). Therefore \mathcal{B} is finite-dimensional.

Conversely if \mathcal{B} is finite-dimensional, the same is true for its restriction to horizontal and vertical lines in \mathbb{Z}^2 . So there exist [3] full rank square matrices $R_i(z_i)$, $i = 1, 2$ such that $\ker R_i(\sigma_i) \supseteq \ker R(\sigma_1, \sigma_2)$. Hence

$$R_i(z_i) = X_i(z_1, z_2)R(z_1, z_2)$$

for some matrices X_i , $i = 1, 2$. By [7] this means that R is factor right prime. \square

An extreme example of nonautonomous systems is the class of (externally) controllable systems. For these systems, the evolution of the trajectories outside a restricted part T of the domain eventually becomes independent of what occurs in T . Formally, we define (external) controllability as follows.

Definition 2. A 2-D system $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ is (externally) controllable if the following condition holds. There exists a positive real number ρ such that, for all $T_1, T_2 \subseteq \mathbb{Z}^2$ with $d(T_1, T_2) > \rho$ and for all $w_1, w_2 \in \mathcal{B}$, there exists $w \in \mathcal{B}$ such that $w|_{T_i} = w_i|_{T_i}$, $i = 1, 2$. Here $d(T_1, T_2)$ denotes the Euclidean distance between the sets T_1 and T_2 .

Remark. We refer to the above notion of controllability as to external controllability in order to make a distinction from the classical notion, which applies to state space realization. Our definition is given at an external level, as it only refers to the (external) system variables $w \in \mathcal{B}$. However, when no possibility of confusion arises, we will simply refer to it as controllability.

An interesting feature of controllability and autonomy is the fact that these are complementary properties, in the sense that an arbitrary AR system can be viewed as the sum of a controllable system with an autonomous one.

Notation. Given two systems $\Sigma_i = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}_i)$, $i = 1, 2$, the sum $\Sigma_1 + \Sigma_2$ of Σ_1 and Σ_2 is defined as $\Sigma_1 + \Sigma_2 := (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$, where $\mathcal{B} := \mathcal{B}_1 + \mathcal{B}_2$.

PROPOSITION 2.3. *Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B})$ be a 2-D system. Then there exist AR systems $\Sigma^c = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}^c)$ and $\Sigma^a = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}^a)$ such that*

1. Σ^c is controllable;
2. Σ^a is autonomous, and
3. $\Sigma = \Sigma^c + \Sigma^a$.

Proof. Let $R(z_1, z_2, z_1^{-1}, z_2^{-1})$ such that

$$\mathcal{B} = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q | R(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0\},$$

Then there exist polynomial matrices $F(z_1, z_2, z_1^{-1}, z_2^{-1})$, with full column rank, and $P(z_1, z_2, z_1^{-1}, z_2^{-1})$, with full row rank and factor left prime, such that $R = FP$. Without loss of generality, we can assume that $P = [P_1 \ P_2]$ with P_1 a square and full rank matrix. Define

$$\mathcal{B}^c = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q | Pw = 0\},$$

$$\mathcal{B}^a = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^q | w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad w_2 = 0, \quad \text{and } FP_1w_1 = 0\},$$

$\Sigma^c = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}^c)$ and $\Sigma^a = (\mathbb{Z}^2, \mathbb{R}^q, \mathcal{B}^a)$. Since P is factor left prime, by [1] Σ^c is controllable. On the other hand, by Proposition 2.2, Σ^a is autonomous. We will show that $\Sigma = \Sigma^c + \Sigma^a$. Since both \mathcal{B}^c and \mathcal{B}^a are subspaces of \mathcal{B} ,

$$\mathcal{B}^c + \mathcal{B}^a \subseteq \mathcal{B}.$$

In order to prove the reciprocal inclusion, assume that $w \in \mathcal{B}$ is given. Let

$$w^a := \begin{bmatrix} w_1^a \\ w_2^a \end{bmatrix},$$

with $w_2^a := 0$ and w_1^a such that $P_1 w_1^a = Pw$ (note that $P_1(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})$ is a surjective operator, as P_1 has full row rank). Further, define $w^c := w - w^a$. Now, since $FP_1 w_1^a = FFW = RW = 0$, it follows that $w^a \in \mathcal{B}^a$. Moreover,

$$Pw^c = Pw - Pw^a = Pw - [P_1 \ P_2] \begin{bmatrix} w_1^a \\ w_2^a \end{bmatrix} = Pw - P_1 w_1^a = 0,$$

and hence $w^c \in \mathcal{B}^c$. So, $w = w^c + w^a$ with $w^c \in \mathcal{B}^c$ and $w^a \in \mathcal{B}^a$, proving that $\mathcal{B} \subseteq \mathcal{B}^c + \mathcal{B}^a$. This yields the desired result. \square

Remark. Note that the above sum $\mathcal{B} = \mathcal{B}^c + \mathcal{B}^a$ (and hence $\Sigma = \Sigma^c + \Sigma^a$) is not necessarily a direct sum, i.e., we may have $\mathcal{B}^c \cap \mathcal{B}^a \neq \{0\}$. However, it can be shown that $\mathcal{B} = \mathcal{B}^c \oplus \mathcal{B}^a$ if and only if, in the decomposition $R = FP$, the matrix P is zero left prime. Moreover, it is not difficult to prove that $\Sigma = \Sigma_1^c + \Sigma_1^a = \Sigma_2^c + \Sigma_2^a$ imply $\Sigma_1^c = \Sigma_2^c$, i.e., in the decomposition $\Sigma = \Sigma^c + \Sigma^a$ the system Σ^c is unique. In fact, it turns out that Σ^c is the largest controllable subsystem of Σ . Curiously, this uniqueness does not necessarily hold for Σ^a . This is illustrated in the following example.

Example. Let $\Sigma = (\mathbb{Z}^2, \mathbb{R}^3, \mathcal{B})$ with

$$\mathcal{B} = \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^3 \mid R(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0\}$$

where

$$R(z_1, z_2, z_1^{-1}, z_2^{-1}) := \begin{bmatrix} z_1 - 1 & 0 & (z_1 - 1)(z_1 + z_2) \\ 0 & z_2 - 1 & (z_2 - 1)(z_2 - z_1) \end{bmatrix}.$$

Then, clearly, R can be decomposed as $R = FP$ with

$$F(z_1, z_2, z_1^{-1}, z_2^{-1}) := \begin{bmatrix} z_1 - 1 & 0 \\ 0 & z_2 - 1 \end{bmatrix}$$

and

$$P(z_1, z_2, z_1^{-1}, z_2^{-1}) := \begin{bmatrix} 1 & 0 & z_1 + z_2 \\ 0 & 1 & z_2 - z_1 \end{bmatrix}.$$

Constructing Σ^c and Σ^a as in the proof of Proposition 2.4 yields $\Sigma^c = (\mathbb{Z}^2, \mathbb{R}^3, \mathcal{B}^c)$ with

$$\mathcal{B}^c := \{w : \mathbb{Z}^2 \rightarrow \mathbb{R}^3 \mid Pw = 0\},$$

and $\Sigma^a = (\mathbb{Z}^2, \mathbb{R}^3, \mathcal{B}^a)$ with

$$\mathcal{B}^a := \left\{ w : \mathbb{Z}^2 \rightarrow \mathbb{R}^3 \mid w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}, w_3 = 0, F \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0 \right\}.$$

So, $\Sigma = \Sigma^a + \Sigma^c$. Now let $\bar{\Sigma}^a = (\mathbb{Z}^2, \mathbb{R}^3, \bar{\mathcal{B}}^a)$, with

$$\bar{\mathcal{B}}^a := \left\{ w : \mathbb{Z}^2 \rightarrow \mathbb{R}^3 \mid w_2 = 0, F\bar{P} \begin{bmatrix} w_1 \\ w_3 \end{bmatrix} = 0 \right\}$$

and

$$\bar{P}(z_1, z_2, z_1^{-1}, z_2^{-1}) := \begin{bmatrix} 1 & z_1 + z_2 \\ 0 & z_1 - z_2 \end{bmatrix}.$$

Applying the same reasoning as in the proof of Proposition 2.3, it is easily shown that, also, $\bar{\Sigma}^a + \Sigma^c = \Sigma$. So, in the decomposition $\Sigma = \Sigma^a + \Sigma^c$ the autonomous subsystem is not unique.

The decomposition of an arbitrary AR 2-D system Σ into the sum of a controllable part Σ^c and an autonomous part Σ^a can be used to obtain state space realizations of Σ by separately realizing Σ^c and Σ^a .

As shown in [1], every controllable AR system admits a state/driving-variable realization of the form

$$\begin{aligned} (2.5) \quad & S(\sigma)\mathbf{x} = 0 \\ (2.6) \quad & \sigma_1\mathbf{x} = (A_1\sigma + A_0)x + (B_1\sigma + B_0)v \\ (2.7) \quad & w = C\mathbf{x} + Dv \end{aligned}$$

with $\sigma := \sigma_2^{-1}\sigma_1$ the diagonal shift, \mathbf{x} the state, and v an auxiliary free driving variable. Moreover, the matrices $S(z), A(z) := A_1z + A_0$ and $B(z) := B_1z + B_0$ are such that $A(\sigma) \ker S(\sigma) \subseteq \ker S(\sigma)$ and $\text{im } B(\sigma) \subseteq \ker S(\sigma)$.

This model can be interpreted as follows. On each diagonal line $\mathcal{L}_k := \{(i, k - i) | i \in \mathbb{Z}\}$, $k \in \mathbb{Z}$, the state trajectories must satisfy the constraint of (2.5). Equation (2.6) yields the state on \mathcal{L}_{k+1} once the state and the driving variable on \mathcal{L}_k are given. We remark that, due to the special structure of $S(z), A(z)$, and $B(z)$, if $\mathbf{x}|_{\mathcal{L}_k}$ satisfies (2.5), then the corresponding state $\mathbf{x}|_{\mathcal{L}_{k+1}}$ computed from (2.6) also satisfies this restriction. Therefore we can view equation (2.5) as a constraint on the admissible initial states along the diagonal line, say \mathcal{L}_0 , and use (2.6) and (2.7) to propagate the (\mathbf{x}, w) trajectories on the half plane $\mathcal{H}_0 := \cup_{k \geq 0} \mathcal{L}_k$. By means of (2.6) the state $\mathbf{x}(i + 1, j)$ is computed from the values of \mathbf{x} and v on the nearest neighbours (i, j) and $(i + 1, j - 1)$ of $(i + 1, j)$.

This updating structure is the same as for the 2-D input-state-output (i/s/o) model introduced in [6], known in the literature as the FM model. However, here the system dynamics are driven by the auxiliary variable v instead of being driven by the system inputs, and the model output is the system variable w , which includes both inputs and outputs. Another important distinction is that the FM i/s/o model does not include an explicit restriction on the admissible initial states as in (2.5). As it will later become clear, such a restriction is essential for the realization of autonomous systems in state/driving-variable form.

In view of foregoing considerations, it turns out that in order to study the realizability of an arbitrary 2-D system $\Sigma = \Sigma^c + \Sigma^a$ by state/driving-variable model as (2.5), (2.6), and (2.7), it is enough to focus on the realizability of the autonomous part Σ^a . This problem will be considered in the next section for the autonomous finite-dimensional case.

3. Autonomous finite-dimensional systems. We will assume throughout that R is a full column rank, factor right prime matrix, with elements in the Laurent polynomial ring $\mathbb{R}[z_1, z_2, z_1^{-1}, z_2^{-1}] := A_{\pm}$. Moreover, for sake of simplicity, we will first restrict to scalar behaviours, by assuming that $R = [r_1 \ r_2 \ \cdots \ r_t]^T$ is a column vector, and successively extend the results to the general case.

3.1. Scalar case. In the subsequent discussion a significant role will be played by some connections between the ideals in A_{\pm} and the ideals in $A_+ := \mathbb{R}[z_1, z_2]$ and by an abstract characterization of the system behaviour based on the algebraic properties of dual spaces. Let us first consider the following map:

$$|\cdot| : A_{\pm} \rightarrow A_+ : p \mapsto |p| := z_1^{-i} z_2^{-j} p,$$

where i and j are the minimum degrees of the monomials that appear in the nonzero Laurent polynomial p with respect to the variables z_1 and z_2 . More precisely, if

$$p = \sum_{h,k \in \mathbb{Z}} p_{hk} z_1^h z_2^k$$

then

$$\begin{aligned} i &:= \min\{h \in \mathbb{Z} \mid \exists k \in \mathbb{Z}, p_{hk} \neq 0\} \\ j &:= \min\{k \in \mathbb{Z} \mid \exists h \in \mathbb{Z}, p_{hk} \neq 0\}. \end{aligned}$$

In case $p = 0$, we define $|p| = 0$. Clearly, for every nonzero Laurent polynomial p , $|p|$ includes a monomial in z_1 and a monomial in z_2 with nonzero coefficients.

The operation just described, of shifting the support of a Laurent polynomial into the positive orthant of $\mathbb{Z} \times \mathbb{Z}$, associates with the ideal $\mathcal{I}_{\pm} := (r_1, r_2, \dots, r_t)_{\pm}$ generated in A_{\pm} by the elements of the matrix R an ideal $\mathcal{I}_+ := (|r_1|, |r_2|, \dots, |r_t|)_+$ generated in A_+ by $|r_1|, |r_2|, \dots, |r_t|$. Some relevant connections between \mathcal{I}_{\pm} and \mathcal{I}_+ are summarized in the following lemma.

LEMMA 3.1. (i) $p \in \mathcal{I}_{\pm}$ if and only if there exists a pair of integers (i, j) such that $z_1^i z_2^j p \in \mathcal{I}_+$.

(ii) The quotient space $A_{\pm}/\mathcal{I}_{\pm}$ is finite-dimensional if and only if the same holds for A_+/\mathcal{I}_+ .

Proof. Statement (i) is obvious. As far as (ii) is concerned, suppose first that A_+/\mathcal{I}_+ is a nonzero finite-dimensional space. This implies that \mathcal{I}_+ , and hence \mathcal{I}_{\pm} , include two nonzero polynomials $f(z_1)$ and $g(z_2)$, with $\deg f > 0$ and $\deg g > 0$. It is easily seen that the cosets

$$[z_1^i z_2^j] := z_1^i z_2^j + \mathcal{I}_{\pm}, \quad 0 \leq i < \deg f, \quad 0 \leq j < \deg g$$

constitute a finite set of generators for the quotient space $A_{\pm}/\mathcal{I}_{\pm}$.

Conversely suppose that $A_{\pm}/\mathcal{I}_{\pm}$ is finite-dimensional and let d be any common factor of $|r_1|, |r_2|, \dots, |r_t|$.

It is clear that $\mathcal{I}_{\pm} \subseteq (d)_{\pm}$, where $(d)_{\pm}$ is the principal ideal of A_{\pm} generated by d . We therefore have $\dim A_{\pm}/\mathcal{I}_{\pm} \geq \dim A_{\pm}/(d)_{\pm}$ and $A_{\pm}/(d)_{\pm}$ is finite-dimensional. This implies that the polynomial sets $\{z_1^i, i \in \mathbb{Z}\}$ and $\{z_2^j, j \in \mathbb{Z}\}$ are linearly dependent modulo $(d)_{\pm}$, and hence $(d)_{\pm}$ includes two nonzero polynomials $f(z_1)$ and $g(z_2)$. Since d must be the constant polynomial, r_1, \dots, r_t are coprime and $A_{\pm}/\mathcal{I}_{\pm}$ is finite-dimensional. \square

We introduce next a special nondegenerate bilinear function

$$\langle \cdot, \cdot \rangle : A_{\pm} \times \mathbb{R}^{\mathbb{Z} \times \mathbb{Z}} \rightarrow \mathbb{R},$$

by assuming

$$\langle p, w \rangle = \sum_{ij} p_{ij} w(i, j)$$

for all polynomials $p = \sum p_{ij} z_1^i z_2^j$ in A_{\pm} and all signals w in $\mathbb{R}^{\mathbb{Z} \times \mathbb{Z}}$.

For instance, if p is the Laurent polynomial $z_1 + z_2^2 - z_1^{-1} z_2 + 3 - z_2^{-3}$ and $w(i, j) = e^{i+j}$, then $\langle p, w \rangle = e^2 + e + 2 - e^{-3}$.

In this way, the “universe” $\mathbb{R}^{\mathbb{Z} \times \mathbb{Z}}$ of all signals with support in $\mathbb{Z} \times \mathbb{Z}$ is isomorphic to the algebraic dual of A_{\pm} , i.e., to the space of the linear functionals on A_{\pm} . Moreover, the

behaviour \mathcal{B} can be identified with the orthogonal complement of \mathcal{I}_\pm with respect to such bilinear function

$$(3.1) \quad \mathcal{B} = \mathcal{I}_\pm^\perp,$$

and, by duality,

$$\mathcal{B}^\perp = \mathcal{I}_\pm^{\perp\perp} = \mathcal{I}_\pm.$$

The proof of (3.1) is an easy consequence of the following identity:

$$p(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w(h, k) = \langle pz_1^h z_2^k, w \rangle.$$

In fact, $w \in \mathcal{B}$ implies $p(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0$ and, therefore, $\langle p, w \rangle = 0, \forall p \in \mathcal{I}_\pm$. Vice versa, given $w \in \mathcal{I}_\pm^\perp$ and $p \in \mathcal{I}_\pm$, we have $\langle pz_1^h z_2^k, w \rangle = 0, \forall h, k \in \mathbb{Z}$, which implies $p(\sigma_1, \sigma_2, \sigma_1^{-1}, \sigma_2^{-1})w = 0$.

We now restrict our attention to the space \mathcal{B} and to the quotient space A_\pm/\mathcal{I}_\pm . Using standard techniques of linear algebra [13], it can be shown that $A_\pm/\mathcal{B}^\perp = A_\pm/\mathcal{I}_\pm$ and \mathcal{B} constitute a dual pair with respect to the bilinear function

$$\langle [p], w \rangle := \langle p, w \rangle.$$

Moreover, the canonical injection

$$i : \mathcal{B} \rightarrow \mathbb{R}^{\mathbb{Z} \times \mathbb{Z}}$$

is dual with respect to the canonical projection π of A_\pm onto A_\pm/\mathcal{I}_\pm

$$\pi : A_\pm/\mathcal{I}_\pm \leftarrow A_\pm.$$

For reasons that will be clear later on, we then wish to exhibit explicitly an isomorphism (for the vector space structure) of \mathcal{B} onto A_\pm/\mathcal{I}_\pm .

PROPOSITION 3.1. *If the matrix R is right factor prime, then \mathcal{B} and A_\pm/\mathcal{I}_\pm are finite-dimensional isomorphic vector spaces.*

Proof. Since R is right factor prime, A_+/\mathcal{I}_+ is finite-dimensional. Therefore, by Lemma 3.1, A_\pm/\mathcal{I}_\pm is finite-dimensional also. Now let $([p_1], [p_2], \dots, [p_n])$ be a basis of A_\pm/\mathcal{I}_\pm and consider the linear map

$$\psi : \mathcal{B} \rightarrow A_\pm/\mathcal{I}_\pm : w \mapsto \sum_{i=1}^n \langle p_i, w \rangle [p_i].$$

If $[\alpha_1 \ \alpha_2 \ \dots \ \alpha_n]^T \in \mathbb{R}^n$ is orthogonal to $[\langle p_1, w \rangle \langle p_2, w \rangle \dots \langle p_n, w \rangle]^T$ for all $w \in \mathcal{B}$, then $\langle \sum \alpha_i p_i, w \rangle = 0$, for all $w \in \mathcal{B}$ and

$$\sum_{i=1}^n \alpha_i p_i \in \mathcal{B}^\perp = \mathcal{I}_\pm.$$

This implies $\alpha_i = 0, i = 1, 2, \dots, n$ and, consequently, as w varies over \mathcal{B} , $[\langle p_1, w \rangle \langle p_2, w \rangle \dots \langle p_n, w \rangle]^T$ span \mathbb{R}^n . Therefore ψ is surjective.

Suppose now that $w \in \mathcal{B}$ satisfies $\psi(w) = 0$ or, equivalently, $\langle p_i, w \rangle = 0, i = 1, 2, \dots, n$. Since every p in A_\pm can be expressed as $p = \sum_{i=1}^n \alpha_i p_i + r, r \in \mathcal{I}_\pm$, for all $p \in A_\pm$ we have $\langle p, w \rangle = \langle r, w \rangle = 0$, which implies $w = 0$. Therefore ψ is injective. \square

From now on we suppose that a basis $([p_1], [p_2], \dots, [p_n])$ has been chosen in $A_{\pm}/\mathcal{I}_{\pm}$, and consider the corresponding dual basis (w_1, w_2, \dots, w_n) in \mathcal{B} .

The relations $\langle [p_i], w_j \rangle = \delta_{ij}$, $i, j = 1, 2, \dots, n$ imply $[p] = \sum_{i=1}^n \langle [p], w_i \rangle [p_i]$, for all $[p] \in A_{\pm}/\mathcal{I}_{\pm}$ and, on the other hand, $w = \sum_{i=1}^n \langle [p_i], w \rangle w_i$, for all $w \in \mathcal{B}$. Introduce the following invertible linear maps:

$$\begin{aligned}\mathcal{Z}_1 : A_{\pm}/\mathcal{I}_{\pm} &\rightarrow A_{\pm}/\mathcal{I}_{\pm} : [p] \mapsto [z_1 p], \\ \mathcal{Z}_2 : A_{\pm}/\mathcal{I}_{\pm} &\rightarrow A_{\pm}/\mathcal{I}_{\pm} : [p] \mapsto [z_2 p].\end{aligned}$$

Clearly, $\mathcal{Z}_1 \mathcal{Z}_2 = \mathcal{Z}_2 \mathcal{Z}_1$ and the adjoint maps of \mathcal{Z}_1 and \mathcal{Z}_2 in \mathcal{B} are σ_1 and σ_2 , respectively.

The matrices $N_l = [n_{hk}^{(l)}]$, $l = 1, 2$, representing the linear transformations \mathcal{Z}_l with respect to the basis $([p_1], [p_2], \dots, [p_n])$ are given by $n_{hk}^{(l)} = \langle [z_l p_k], w_h \rangle$. Hence, the matrices representing σ_1 and σ_2 with respect to the dual basis are N_1^T and N_2^T , respectively. In fact, letting $\sigma_l w_j = \sum_h t_{hj}^{(l)} w_h$, $l = 1, 2$, we have

$$(3.2) \quad \langle [p_k], \sigma_l w_j \rangle = \sum_h t_{hj}^{(l)} \langle [p_k], w_h \rangle = t_{kj}^{(l)}$$

and, using the duality,

$$(3.3) \quad \langle [p_k], \sigma_l w_j \rangle = \langle [z_l p_k], w_j \rangle = \sum_r n_{rk}^{(l)} \langle p_r, w_j \rangle = n_{jk}^{(l)}.$$

Comparing (3.2) and (3.3) gives the result.

We are now in a position to provide a state driving-variable realization of an autonomous finite-dimensional system Σ^a .

For any $w \in \mathcal{B}$, we introduce the following signal:

$$\mathbf{x} : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{R}^n : (h, k) \mapsto \begin{bmatrix} \langle [p_1], \sigma_1^h \sigma_2^k w \rangle \\ \langle [p_2], \sigma_1^h \sigma_2^k w \rangle \\ \vdots \\ \langle [p_n], \sigma_1^h \sigma_2^k w \rangle \end{bmatrix}.$$

The value of \mathbf{x} at (h, k) provides the components of $\sigma_1^h \sigma_2^k w$ with respect to the basis (w_1, w_2, \dots, w_n) . It is clear that, once $\mathbf{x}(0, 0)$ is known, $\mathbf{x}(h, k)$ is easily computed for all $(h, k) \in \mathbb{Z} \times \mathbb{Z}$

$$\mathbf{x}(h, k) = (N_1^T)^h (N_2^T)^k \mathbf{x}(0, 0).$$

Moreover, the value of w at (h, k) can be recovered from $\mathbf{x}(h, k)$ as follows:

$$\begin{aligned}w(h, k) &= (\sigma_1^h \sigma_2^k w)(0, 0) = \langle [1], \sigma_1^h \sigma_2^k w \rangle \\ &= \langle c_1 [p_1] + c_2 [p_2] + \dots + c_n [p_n], \sigma_1^h \sigma_2^k w \rangle = C \mathbf{x}(h, k).\end{aligned}$$

Here $C := [c_1 \ c_2 \ \dots \ c_n]$ denotes the row vector of the components of $[1]$ with respect to the basis $([p_1], [p_2], \dots, [p_n])$ in $A_{\pm}/\mathcal{I}_{\pm}$.

The above results are summarized in the following recursive model:

$$\begin{cases} \sigma_1 \mathbf{x} = N_1^T \mathbf{x} \\ \sigma_2 \mathbf{x} = N_2^T \mathbf{x} \\ w = C \mathbf{x}. \end{cases}$$

Every signal of the autonomous behaviour \mathcal{B} is uniquely determined by the corresponding value of the state \mathbf{x} at any point (h, k) and, conversely, different states at (h, k) induce different

signals in the autonomous system. Finally, letting $\sigma := \sigma_1 \sigma_2^{-1}$, $S(\sigma) := \sigma I - N_1^T (N_2^T)^{-1}$, and $A(\sigma) = N_1^T$, we end up with a state driving-variable realization of \mathcal{B} , as follows:

$$\begin{cases} S(\sigma)w = 0 \\ \sigma_1 \mathbf{x} = A(\sigma)\mathbf{x} \\ w = C\mathbf{x}. \end{cases}$$

When $p_j, j = 1, 2, \dots, n$ are monic monomials, i.e., $p_j = z_1^{\mu_j} z_2^{\nu_j}, j = 1, 2, \dots, n$, the structure of the corresponding dual basis is very appealing. In fact, the element w_j is the unique element of \mathcal{B} taking the values 1 at (μ_j, ν_j) and 0 at $(\mu_i, \nu_i), i = 1, 2, \dots, j-1, j+1, \dots, n$. Moreover, for every $(h, k) \in \mathbb{Z} \times \mathbb{Z}$, the components of the state vector $\mathbf{x}(h, k)$ are the values of w at $\{(\mu_1 + h, \nu_1 + k), (\mu_2 + h, \nu_2 + k), \dots, (\mu_n + h, \nu_n + k)\}$.

A further reason for using a monomial basis in $\mathcal{A}_\pm/\mathcal{I}_\pm$ will be made apparent in the following subsection, where an algorithm for the computation of the matrices N_1^T and N_2^T is outlined. Some concepts on computer algebra, and in particular on Gröbner basis theory, are required; for details see [10].

3.2. Computational methods. Let $\mathcal{G} = \{g_1, g_2, \dots, g_h\}, g_i \in A_+$, be a Gröbner basis of the ideal \mathcal{I}_+ , and denote by $\{q_1 = 1, q_2, \dots, q_m\}$ the set of monic monomials that are not multiple of the leading power products of any of the polynomials in \mathcal{G} . Then

$$(q_1 + \mathcal{I}_+, q_2 + \mathcal{I}_+, \dots, q_m + \mathcal{I}_+)$$

is a basis of A_+/\mathcal{I}_+ and the linear transformations

$$\phi_l : A_+/\mathcal{I}_+ \rightarrow A_+/\mathcal{I}_+ : q + \mathcal{I}_+ \mapsto z_l q + \mathcal{I}_+, \quad l = 1, 2$$

are represented by a pair of commuting matrices M_1 and M_2 .

Our purpose here is to supplement the algorithm discussed in [11] for obtaining M_1 and M_2 , so as to provide a constructive technique for obtaining the invertible matrices N_1 and N_2 introduced in the previous subsection. The procedure we are going to describe will also shed some light on the connections between the ideals \mathcal{I}_\pm and \mathcal{I}_+ .

Let μ be a positive integer with the property that the subspace of A_+/\mathcal{I}_+ spanned by $\{z_1^{\mu+h} z_2^{\mu+k} q_i + \mathcal{I}_+, i = 1, 2, \dots, m\}$ is independent of h and k , for all h and $k \geq 0$. Therefore $\mathcal{L} := \text{span}\{z_1^\mu z_2^\mu q_i + \mathcal{I}_+, i = 1, 2, \dots, m\}$ is a ϕ_1 - ϕ_2 -invariant subspace, satisfying $\phi_1 \mathcal{L} = \phi_2 \mathcal{L} = \mathcal{L}$ and the restrictions of ϕ_1 and ϕ_2 to \mathcal{L} constitute a couple of invertible commutative linear transformations.

Let S be the Boolean matrix that selects a basis of \mathcal{L} out of the ordered array $(z_1^\mu z_2^\mu q_1 + \mathcal{I}_+, z_1^\mu z_2^\mu q_2 + \mathcal{I}_+, \dots, z_1^\mu z_2^\mu q_m + \mathcal{I}_+)$

$$(3.4) \quad \begin{aligned} & (z_1^\mu z_2^\mu q_1 + \mathcal{I}_+, z_1^\mu z_2^\mu q_2 + \mathcal{I}_+, \dots, z_1^\mu z_2^\mu q_m + \mathcal{I}_+) S \\ & = (z_1^\mu z_2^\mu q_{i_1} + \mathcal{I}_+, z_1^\mu z_2^\mu q_{i_2} + \mathcal{I}_+, \dots, z_1^\mu z_2^\mu q_{i_\nu} + \mathcal{I}_+). \end{aligned}$$

So the restriction $\phi_1|_{\mathcal{L}}$ is associated with a $\nu \times \nu$ invertible matrix N_1 , and, recalling that M_1 and M_2 represent the linear transformations ϕ_1 and ϕ_2 with respect to the basis $(q_1 + \mathcal{I}_+, q_2 + \mathcal{I}_+, \dots, q_m + \mathcal{I}_+)$, we have

$$(3.5) \quad \begin{aligned} & (z_1^{\mu+1} z_2^\mu q_{i_1} + \mathcal{I}_+, z_1^{\mu+1} z_2^\mu q_{i_2} + \mathcal{I}_+, \dots, z_1^{\mu+1} z_2^\mu q_{i_\nu} + \mathcal{I}_+) \\ & = (q_1 + \mathcal{I}_+, q_2 + \mathcal{I}_+, \dots, q_m + \mathcal{I}_+) M_1^\mu M_2^\mu S N_1. \end{aligned}$$

On the other hand, the definitions of M_1 , M_2 , and S also imply

$$(3.6) \quad \begin{aligned} & (z_1^{\mu+1} z_2^\mu q_{i_1} + \mathcal{I}_+, z_1^{\mu+1} z_2^\mu q_{i_2} + \mathcal{I}_+, \dots, z_1^{\mu+1} z_2^\mu q_{i_\nu} + \mathcal{I}_+) \\ & = (q_1 + \mathcal{I}_+, q_2 + \mathcal{I}_+, \dots, q_m + \mathcal{I}_+) M_1^{\mu+1} M_2^\mu S. \end{aligned}$$

Comparing (3.5) and (3.6) gives $M_1^{\mu+1}M_2^\mu S = M_1^\mu M_2^\mu S N_1$. Since $M_1^\mu M_2^\mu S$ has full column rank, letting

$$H := (M_1^T)^\mu (M_2^T)^\mu M_1^\mu M_2^\mu,$$

we obtain

$$(3.7) \quad N_1 = (S^T H S)^{-1} (S^T H M_1 S).$$

Similarly,

$$(3.8) \quad N_2 = (S^T H S)^{-1} (S^T H M_2 S).$$

The next proposition shows that the monomials $q_{i_1}, q_{i_2}, \dots, q_{i_\nu}$ resulting from the previous procedure and associated, as shown, to a basis of the subspace $\mathcal{L} \subseteq A_+/\mathcal{I}_+$, also provide the basis of the quotient space A_\pm/\mathcal{I}_\pm we are looking for.

PROPOSITION 3.2. *The monomials $q_{i_1}, q_{i_2}, \dots, q_{i_\nu}$ constitute a basis of A_\pm , modulo \mathcal{I}_\pm .*

Proof. Suppose that $\sum_{h=1}^n \alpha_h q_{i_h}$ is in \mathcal{I}_\pm .

By Lemma 3.1 there exists a positive integer λ such that $\sum_{h=1}^\nu \alpha_h q_{i_h} z_1^\lambda z_2^\lambda$ and, a fortiori, $\sum_{h=1}^\nu \alpha_h q_{i_h} z_1^{\mu+\lambda} z_2^{\mu+\lambda}$ belong to \mathcal{I}_+ .

Since the monomials $q_{i_1} z_1^{\mu+\lambda} z_2^{\mu+\lambda}, q_{i_2} z_1^{\mu+\lambda} z_2^{\mu+\lambda}, \dots, q_{i_\nu} z_1^{\mu+\lambda} z_2^{\mu+\lambda}$ are linearly independent modulo \mathcal{I}_+ , we have $\alpha_h = 0, h = 1, 2, \dots, \nu$, and $q_{i_h}, h = 1, 2, \dots, \nu$ are linearly independent modulo \mathcal{I}_\pm .

It remains to show that they generate A_\pm modulo \mathcal{I}_\pm . To that purpose, consider any polynomial $p \in A_\pm$. Then there exists a positive integer λ such that $z_1^\lambda z_2^\lambda p \in A_+$. Therefore

$$\begin{aligned} (z_1^\lambda z_2^\lambda p) z_1^\mu z_2^\mu &= \sum_{h=1}^\nu \alpha_h q_{i_h} z_1^\mu z_2^\mu = \sum_{h=1}^\nu \beta_h q_{i_h} z_1^{\mu+\lambda} z_2^{\mu+\lambda} \bmod \mathcal{I}_+ \\ &= \sum_{h=1}^\nu \beta_h q_{i_h} z_1^{\mu+\lambda} z_2^{\mu+\lambda} \bmod \mathcal{I}_\pm. \end{aligned}$$

Upon multiplying on both sides by $z_1^{-\mu-\lambda} z_2^{-\mu-\lambda}$, we have $p = \sum_{h=1}^\nu \beta_h q_{i_h} \bmod \mathcal{I}_\pm$, showing that the monomials q_{i_h} generate A_\pm modulo \mathcal{I}_\pm . \square

As a consequence of Proposition 3.2, the matrices N_1 and N_2 associated with the restrictions to \mathcal{L} of ϕ_1 and ϕ_2 , with respect to the basis (3.4), represent \mathcal{Z}_1 and \mathcal{Z}_2 with respect to the basis

$$(3.9) \quad ([q_{i_1}], [q_{i_2}], \dots, [q_{i_\nu}])$$

in A_\pm/\mathcal{I}_\pm . This result is almost obvious. Upon introducing the following isomorphism $\psi : \mathcal{L} \rightarrow A_\pm/\mathcal{I}_\pm : \sum_{h=1}^\nu \alpha_h (z_1^\mu z_2^\mu q_{i_h}) + \mathcal{I}_+ \mapsto \sum_{h=1}^\nu \alpha_h [q_{i_h}]$, we check that the following diagram

$$\begin{array}{ccc} \mathcal{L} & \xrightarrow{\psi} & A_\pm/\mathcal{I}_\pm \\ \downarrow \phi_1 & & \downarrow \mathcal{Z}_1 \\ \mathcal{L} & \xrightarrow{\psi} & A_\pm/\mathcal{I}_\pm \end{array}$$

commutes. With respect to the bases (3.4) and (3.9), ψ is represented by the identity matrix and therefore the same matrix N_1 represents both ϕ_1 and \mathcal{Z}_1 . Similarly ϕ_2 and \mathcal{Z}_2 are both represented by N_2 .

Remark. In [11] it has been shown that the annihilating polynomials of M_1 and M_2 are exactly the polynomials of the ideal \mathcal{I}_+ , i.e.,

$$p(M_1, M_2) = 0 \Leftrightarrow p \in \mathcal{I}_+.$$

It is quite natural to ask whether the Laurent polynomials in \mathcal{I}_\pm do exhibit the characteristic property of annihilating the commutative invertible matrices N_1 and N_2 . Actually this is true and we have

$$(3.10) \quad p(N_1, N_2) = 0 \Leftrightarrow p \in \mathcal{I}_\pm.$$

To prove (3.10), we note first that, by Lemma 3.1, $p \in \mathcal{I}_\pm$ if and only if there exists a pair of nonnegative integers i and j , such that $q(z_1, z_2) := z_1^i z_2^j p(z_1, z_2) \in \mathcal{I}_+$. This in turn is equivalent to assuming that $0 = q(M_1, M_2) = q(\phi_1, \phi_2)$ and therefore (3.10) can be restated as follows:

$$(3.11) \quad p(N_1, N_2) = 0 \Leftrightarrow q(M_1, M_2) = 0$$

for some $q = z_1^i z_2^j p \in A_\pm$. To prove (3.11), assume first $q(M_1, M_2) = 0$. Then $q(\phi_1, \phi_2) = 0$ implies $q(\phi_1|_{\mathcal{L}}, \phi_2|_{\mathcal{L}}) = 0$ and, consequently, $0 = q(N_1, N_2) = N_1^i N_2^j p(N_1, N_2) = p(N_1, N_2)$, because of the invertibility of N_1 and N_2 .

Vice versa, consider any polynomial $p \in A_\pm$ that annihilates the commutative pair N_1, N_2 , i.e.,

$$(3.12) \quad p(N_1, N_2) = 0.$$

Select a pair of nonnegative integers h, k such that $p' = z_1^h z_2^k p$ is a polynomial in A_+ . Rewrite p' as follows:

$$p' = \sum_{j=1}^m \beta_j q_j + r, \quad r \in \mathcal{I}_+$$

and let

$$q := z_1^\mu z_2^\mu p' = \sum_{j=1}^m \beta_j z_1^\mu z_2^\mu q_j + z_1^\mu z_2^\mu r.$$

Note that (3.12) implies $p'(N_1, N_2) = q(N_1, N_2) = 0$ and $r \in \mathcal{I}_+$ implies $r(M_1, M_2) = 0$. Restricting ϕ_1 and ϕ_2 to \mathcal{L} gives $r(N_1, N_2) = 0$ and hence $\sum_{j=1}^m \beta_j q_j(N_1, N_2) = 0$. To prove that $q(M_1, M_2)$ is the zero matrix, we will show that $q(\phi_1, \phi_2)$ annihilates $q_i + \mathcal{I}_\pm$, $i = 1, 2, \dots, m$. Actually we have

$$\begin{aligned} q(\phi_1, \phi_2)(q_i + \mathcal{I}_+) &= \sum_{j=1}^m \beta_j \phi_1^\mu \phi_2^\mu q_j(\phi_1, \phi_2)(q_i + \mathcal{I}_+) + \phi_1^\mu \phi_2^\mu r(\phi_1, \phi_2)(q_i + \mathcal{I}_+) \\ &= \sum_{j=1}^m \beta_j q_j(\phi_1, \phi_2)(z_1^\mu z_2^\mu q_i + \mathcal{I}_+). \end{aligned}$$

Since $z_1^\mu z_2^\mu q_i + \mathcal{I}_+$, $i = 1, 2, \dots, m$ belong to \mathcal{L} and $\sum_j \beta_j q_j(\phi_1, \phi_2)$ acts on \mathcal{L} as the zero transformation, we are done.

Remark. An alternative way for computing the matrices N_1 and N_2 can be derived from the scheme presented in [3]. Actually, when \mathcal{B} is finite-dimensional that scheme gives a finite set of initial conditions that can be viewed as the state vector of a state space realization. The updating matrices in that case are obtained from the canonical computational form proposed there.

3.3. Extension to the vector case. Suppose now that R is a $t \times q$ full column rank right prime matrix, describing a q variables behaviour \mathcal{B} . All concepts previously introduced for the scalar case have an immediate extension to the vector case. Let

$$A_+^q := \mathbb{R}^{1 \times q}[z_1, z_2]$$

$$A_\pm^q := \mathbb{R}^{1 \times q}[z_1, z_2, z_1^{-1}, z_2^{-1}]$$

and define the map

$$|\cdot| : A_\pm^q \rightarrow A_+^q : r \mapsto |r| := z_1^i z_2^j r,$$

where i and j are the minimum degrees of r with respect to the indeterminates z_1 and z_2 . In case $p = 0$, we define $|p| = 0$.

Let $\mathcal{M}_\pm := (r_1, \dots, r_t)_\pm$ be the submodule in A_\pm^q generated by the rows of R and $\mathcal{M}_+ := (|r_1|, \dots, |r_t|)_+$ the submodule in A_+^q generated by the rows of the matrix

$$\overline{R} := \begin{bmatrix} |r_1| \\ \vdots \\ |r_t| \end{bmatrix} = \Lambda R$$

where $\Lambda := \text{diag} \{z_1^{\nu_1} z_2^{\mu_1}, \dots, z_1^{\nu_t} z_2^{\mu_t}\}$ and ν_i and μ_i satisfy $z_1^{\nu_i} z_2^{\mu_i} r_i = |r_i|$, $i = 1, \dots, t$.

LEMMA 3.2. (i) A row r belongs to \mathcal{M}_\pm if and only if there exists a pair of integers (i, j) such that $z_1^i z_2^j r$ is in \mathcal{M}_+ .

(ii) $A_\pm^q / \mathcal{M}_\pm$ is finite-dimensional if and only if A_+^q / \mathcal{M}_+ is finite-dimensional.

Proof. (i) The proof is obvious.

(ii) Suppose that A_+^q / \mathcal{M}_+ is finite-dimensional. This implies that there exist polynomials $f_i(z_1)$ and $g_i(z_2)$, $i = 1, \dots, q$, such that

$$f_i(z_1)e_i^T \in \mathcal{M}_+$$

$$g_i(z_2)e_i^T \in \mathcal{M}_+$$

where e_i is the element of the canonical basis of \mathbb{R}^q with one in position i . It is easily seen that

$$\bigcup_{i=1}^q \{z_1^h z_2^k e_i^T + \mathcal{M}_+, 0 \leq h < \deg f_i, 0 \leq k < \deg g_i\}$$

constitutes a set of generators for $A_\pm^q / \mathcal{M}_\pm$.

Conversely, suppose that $A_\pm^q / \mathcal{M}_\pm$ is finite dimensional and let D be any right factor of \overline{R} ,

$$\overline{R} = \hat{R}D.$$

It follows that $R = \Lambda^{-1} \overline{R} = \Lambda^{-1} \hat{R}D$, where Λ^{-1} is still a polynomial matrix with elements in A_\pm . Since the submodule $\mathcal{M}_\pm(D)$ generated by the rows of D , satisfies $\mathcal{M}_\pm \subseteq \mathcal{M}_\pm(D)$, we have $\dim A_\pm^q / \mathcal{M}_\pm \geq \dim A_\pm^q / \mathcal{M}_\pm(D)$ and $A_\pm^q / \mathcal{M}_\pm(D)$ is finite-dimensional. Therefore there exist polynomials $f_i(z_1)$ and $g_i(z_2)$, $i = 1, \dots, q$, such that

$$f_i(z_1)e_i^T \in \mathcal{M}_\pm(D)$$

$$g_i(z_2)e_i^T \in \mathcal{M}_\pm(D)$$

and, consequently, there exist polynomial matrices H and K such that

$$(3.13) \quad HD = \text{diag}\{f_1(z_1), \dots, f_q(z_1)\}$$

$$(3.14) \quad KD = \text{diag}\{g_1(z_2), \dots, g_q(z_2)\}.$$

Equation (3.13) implies that $\det D$ is a polynomial in z_1 and (3.14) implies that $\det D$ is a polynomial in z_2 . Therefore D is unimodular, \bar{R} is right factor prime and A_+^q/\mathcal{M}_+ is finite dimensional. \square

Introduce a nondegenerate bilinear function

$$\langle \cdot, \cdot \rangle : A_{\pm}^q \times (\mathbb{R}^q)^{\mathbb{Z} \times \mathbb{Z}} \rightarrow \mathbb{R},$$

such that $\langle r, w \rangle = \sum r_{ij}w(i, j)$, where $r = \sum r_{ij}z_1^i z_2^j$ is a polynomial row in A_{\pm}^q and $w \in (\mathbb{R}^q)^{\mathbb{Z} \times \mathbb{Z}}$.

Then $(\mathbb{R}^q)^{\mathbb{Z} \times \mathbb{Z}}$ is isomorphic to the algebraic dual of A_{\pm}^q and we still have $\mathcal{B} = \mathcal{M}_{\pm}^{\perp}$ and $\mathcal{B}^{\perp} = \mathcal{M}_{\pm}^{\perp\perp} = \mathcal{M}_{\pm}$. Moreover \mathcal{B} and $A_{\pm}^q/\mathcal{M}_{\pm}$ are finite-dimensional isomorphic vector spaces.

As in the scalar case, let N_1 and N_2 be the matrices of the linear transformations

$$\begin{aligned} \mathcal{Z}_1 : A_{\pm}^q/\mathcal{M}_{\pm} &\rightarrow A_{\pm}^q/\mathcal{M}_{\pm} : [r] \mapsto [z_1 r] \\ \mathcal{Z}_2 : A_{\pm}^q/\mathcal{M}_{\pm} &\rightarrow A_{\pm}^q/\mathcal{M}_{\pm} : [r] \mapsto [z_2 r] \end{aligned}$$

with respect to the basis $([r_1], \dots, [r_n])$ of $A_{\pm}^q/\mathcal{M}_{\pm}$. If the state vector relative to a signal $w \in \mathcal{B}$ is defined as

$$\mathbf{x}(h, k) := \begin{bmatrix} \langle \mathcal{Z}_1^h \mathcal{Z}_2^k [r_1], w \rangle \\ \vdots \\ \langle \mathcal{Z}_1^h \mathcal{Z}_2^k [r_n], w \rangle \end{bmatrix},$$

and C is a $q \times n$ constant matrix such that

$$\begin{bmatrix} [e_1^T] \\ \vdots \\ [e_q^T] \end{bmatrix} = C \begin{bmatrix} [r_1] \\ \vdots \\ [r_n] \end{bmatrix},$$

then we have

$$\mathbf{x}(h, k) := (N_1^T)^h (N_2^T)^k \mathbf{x}(0, 0)$$

and

$$w(h, k) = C\mathbf{x}(h, k).$$

By applying the theory of Gröbner basis over the polynomial modules [12], a basis in A_+^q/\mathcal{M}_+ with elements of the type $z_1^h z_2^k e_i^T + \mathcal{M}_+$ is easily obtained. After computing the matrices M_1 and M_2 that represent the transformations

$$\phi_l : A_+^q/\mathcal{M}_+ \rightarrow A_+^q/\mathcal{M}_+ : r + \mathcal{M}_+ \mapsto z_l r + \mathcal{M}_+, \quad l = 1, 2$$

with respect to that basis, the procedure for extracting N_1 and N_2 from M_1 and M_2 is the same introduced in the scalar case.

REFERENCES

- [1] P. ROCHA AND J. C. WILLEMS, *Controllability of 2-D systems*, IEEE Trans. on Automatic Control, AC-36 (1991), pp. 413–423.
- [2] J. C. WILLEMS, *Models for dynamics*, Dynamics Reported, 2 (1989), pp. 171–269.
- [3] P. ROCHA AND J. C. WILLEMS, *Canonical computational forms for AR 2-D systems*, Multidim. Systems Signal Processing, 1 (1990), pp. 251–278.
- [4] S. ATTASI, *Systèmes linéaires homogènes à deux indices*, Rapport Laboria, 31 (1973).
- [5] R. P. ROESSER, *A discrete state space model for linear image processing*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 1–10.
- [6] E. FORNASINI AND G. MARCHESINI, *State space realization theory of 2-D filters*, IEEE Trans. Automatic. Control, AC-21 (1976), pp. 484–492.
- [7] M. MORF, B. C. LEVY, AND S. Y. KUNG, *New results in 2-D systems theory*, Part I, Proceedings IEEE, 65 (1977), pp. 861–872.
- [8] T. KACZOREK, *Singular general model of 2-D systems and its solutions*, IEEE Trans. Automat. Control, AC-31 (1988), pp. 1060–1061.
- [9] F. L. LEWIS AND B. G. MERTIOS, *On the analysis of 2-D discrete singular systems*, preprint.
- [10] B. BUCHBERGER, *Gröbner basis: An algorithmic method in polynomial ideal theory*, Multidimensional Systems Theory, N. K. Bose, ed., D. Reidel, Boston, MA, pp. 184–232.
- [11] E. FORNASINI, *A note on output feedback stabilizability of multivariable 2-D systems*, Systems Control Lett., 10 (1988), pp. 45–50.
- [12] H. M. MÖLLER AND F. MORA, *New constructive methods in classical ideal theory*, J. Algebra, 100 (1986), pp. 138–178.
- [13] W. GREUB, *Linear Algebra*, Springer-Verlag, New York, 1975.

ROBUST STABILITY OF FEEDBACK SYSTEMS: A GEOMETRIC APPROACH USING THE GAP METRIC*

CIPRIAN FOIAS[†], TRYPHON T. GEORGIU[‡], AND MALCOLM C. SMITH[§]

Abstract. A geometric framework for robust stabilization of infinite-dimensional time-varying linear systems is presented. The uncertainty of a system is described by perturbations of its graph and is measured in the gap metric. Necessary and sufficient conditions for robust stability are generalized from the time-invariant case. An example is given to highlight an important difference between the obstructions, which limit the size of a stabilizable gap ball, in the time-varying and time-invariant cases. Several results on the gap metric and the gap topology are established that are central in a geometric treatment of the robust stabilizability problem in the gap. In particular, the concept of a “graphable” subspace is introduced in the paper. Subspaces that fail to be graphable are characterized by an index condition on a certain semi-Fredholm operator.

Key words. Robust stabilization, gap metric, graph topology, graphability, stabilizability

AMS subject classifications. 47A53, 47N70, 93B27, 93B28, 93B36, 93C25, 93C50, 93D25

1. Introduction. In this paper we develop a geometric framework for robust stabilization of feedback systems using operator-theoretic methods. The theory is based on a description of the uncertainty of a system as a perturbation of its graph and is measured by the gap metric.

The gap metric has its origin in functional analysis [20], [13], where it was used in perturbation theory of linear operators. It was introduced into control theory in [25], [1] as being appropriate for the study of uncertainty in feedback systems. For shift-invariant systems it was shown in [5] that the gap metric was computable exactly in terms of two standard “2-block” H_∞ optimization problems. Building on this result and the work of [23], [24], [26], and [7], it was shown in [6] that robust stabilization in the gap metric is equivalent to robust stabilization for perturbations of the normalized coprime factors of the transfer function.

The simplicity of the robustness bounds obtained in [6] for the time-invariant case, which were expressed solely in terms of the plant and controller system operators, strongly suggests potential generalization. However, the techniques used in [6] are mostly function theoretic, relying on a specific representation for the graph of a time-invariant dynamical system as a shift-invariant subspace of $L_2[0, \infty]$, and do not admit immediate generalization to the shift-varying case. This motivated the search for a different approach, which does not rely on representations for the subspaces involved, and which elucidates the apparent geometric structure underlying the robust stabilization problem. It became apparent that substantially new techniques were needed, beyond those developed in [6], to meet this objective. The present paper is a continuation of work begun in [3], [4]. We note that some independent work on a geometric approach to robust stability in the gap metric has been presented in [15], [16], [19]. A generalization of the results of [5] has been presented in [2].

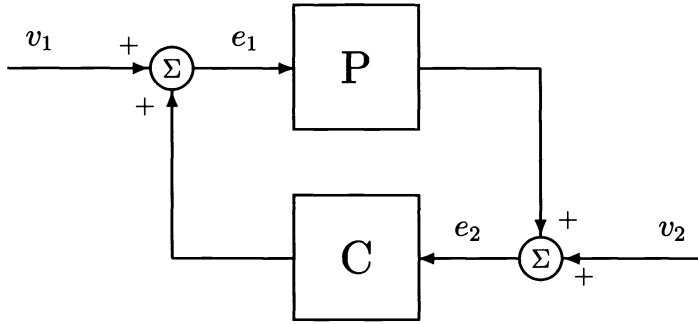
This paper is organized as follows. In §2 we present some basic material on

* Received by the editors September 30, 1991; accepted for publication (in revised form) August 17, 1992. This work was supported in part by National Science Foundation grants DMS-8802596, ECS-9016050, and INT-9024869, Science and Engineering Research Council grant GE/H15653, the OIE, and the Graduate School at the University of Minnesota.

[†] Department of Mathematics, University of Indiana, Bloomington, Indiana 47405.

[‡] Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

[§] Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom.

FIG. 1. *Standard feedback configuration*

graphs and stabilizability for linear systems. In §3 we establish several results on the gap metric that are used in the later development. Section 4 introduces the concept of graphability and proves a necessary and sufficient condition for a subspace to be graphable. Section 5 presents and proves the main robustness theorem for plant uncertainty in the gap metric. In §6 an example is presented to clarify the need for the uniform boundedness condition in the main robustness theorem. Section 7 uses the machinery of the previous sections to generalize an elegant result of Qiu and Davison [16] on combined plant-controller uncertainty to the time-varying case.

2. Graphs and stabilizability of linear systems. We consider a linear system to be a (possibly unbounded) linear operator $\mathbf{P} : \mathcal{D}_{\mathbf{P}} \subset \mathcal{U} \rightarrow \mathcal{Y}$, where \mathcal{U}, \mathcal{Y} are Hilbert spaces and $\mathcal{D}_{\mathbf{P}}$ is the *domain* of \mathbf{P} . We denote by $\mathcal{P}_{\mathcal{U}, \mathcal{Y}}$ the class of all linear systems from \mathcal{U} to \mathcal{Y} . A typical choice for the input and output spaces is $\mathcal{U} = \ell_2^m[0, \infty)$ and $\mathcal{Y} = \ell_2^p[0, \infty)$, or the corresponding continuous-time Lebesgue spaces. (Note: This paper does not impose the constraints of causality or time-invariance on the systems considered.)

Consider the feedback configuration of Fig. 1, where the *plant* $\mathbf{P} \in \mathcal{P}_{\mathcal{U}, \mathcal{Y}}$ and the *controller* $\mathbf{C} \in \mathcal{P}_{\mathcal{Y}, \mathcal{U}}$. This configuration, denoted by $[\mathbf{P}, \mathbf{C}]$, provides a pictorial representation of the following set of equations:

$$\begin{aligned} e_1 &= v_1 + Ce_2, \\ e_2 &= Pe_1 + v_2. \end{aligned}$$

Define the *graph* of a system $\mathbf{P} \in \mathcal{P}_{\mathcal{U}, \mathcal{Y}}$ as the linear manifold of bounded input-output pairs of \mathbf{P}

$$\mathcal{G}_{\mathbf{P}} := \begin{pmatrix} \mathbf{I}_{\mathcal{U}} \\ \mathbf{P} \end{pmatrix} \mathcal{D}_{\mathbf{P}} \subset \mathcal{L} := \mathcal{U} \oplus \mathcal{Y},$$

where $\mathbf{I}_{\mathcal{U}}$ denotes the identity operator on \mathcal{U} . Similarly, define the *inverse graph* of the controller \mathbf{C} by

$$\mathcal{G}'_{\mathbf{C}} := \begin{pmatrix} \mathbf{C} \\ \mathbf{I}_{\mathcal{Y}} \end{pmatrix} \mathcal{D}_{\mathbf{C}} \subset \mathcal{L}.$$

The feedback configuration $[\mathbf{P}, \mathbf{C}]$ is said to be *stable* if the operators mapping $v_i \rightarrow e_j$ for $i, j = 1, 2$ are bounded. This is equivalent to the operator

$$\mathbf{F}_{\mathbf{P}, \mathbf{C}} := \begin{pmatrix} \mathbf{I}_{\mathcal{U}} & \mathbf{C} \\ \mathbf{P} & \mathbf{I}_{\mathcal{Y}} \end{pmatrix} : \mathcal{D}_{\mathbf{P}} \times \mathcal{D}_{\mathbf{C}} \rightarrow \mathcal{G}_{\mathbf{P}} + \mathcal{G}'_{\mathbf{C}} : \begin{pmatrix} e_1 \\ -e_2 \end{pmatrix} \rightarrow \begin{pmatrix} v_1 \\ -v_2 \end{pmatrix}$$

having a bounded inverse defined on \mathcal{L} . In case an inverse exists it is denoted by $\mathbf{H}_{\mathbf{P}, \mathbf{C}} := \mathbf{F}_{\mathbf{P}, \mathbf{C}}^{-1}$.

A system \mathbf{P} is said to be *stabilizable* if and only if there exists a controller \mathbf{C} such that $[\mathbf{P}, \mathbf{C}]$ is stable.

PROPOSITION 1. *If $\mathbf{P} \in \mathcal{P}_{\mathcal{U}, \mathcal{Y}}$ is stabilizable then the graph of \mathbf{P} is closed.*

Proof. Let $\begin{pmatrix} u_i \\ y_i \end{pmatrix} \in \mathcal{G}_{\mathbf{P}}$ for $i = 1, 2, \dots$ be a Cauchy sequence with limit point $\begin{pmatrix} u \\ y \end{pmatrix}$, and let \mathbf{C} be such that $[\mathbf{P}, \mathbf{C}]$ is stable. Since

$$\begin{pmatrix} u_i \\ 0 \end{pmatrix} := \mathbf{H}_{\mathbf{P}, \mathbf{C}} \begin{pmatrix} u_i \\ y_i \end{pmatrix},$$

then

$$\begin{pmatrix} u \\ 0 \end{pmatrix} := \lim_{i \rightarrow \infty} \mathbf{H}_{\mathbf{P}, \mathbf{C}} \begin{pmatrix} u_i \\ y_i \end{pmatrix} = \mathbf{H}_{\mathbf{P}, \mathbf{C}} \begin{pmatrix} u \\ y \end{pmatrix},$$

which means that $u \in \mathcal{D}_{\mathbf{P}}$ and $\begin{pmatrix} u \\ y \end{pmatrix} = (\mathbf{I}_{\mathbf{P}})u \in \mathcal{G}_{\mathbf{P}}$. \square

Thus, a necessary condition for $[\mathbf{P}, \mathbf{C}]$ to be stable is that both \mathbf{P} and \mathbf{C} have closed graphs. A similar statement has been made in [27] for quotients of bounded operators. The idea in [27] gives Proposition 1 in the following way: Closed-loop stability ensures that $\mathbf{P}(\mathbf{I} + \mathbf{C}\mathbf{P})^{-1}$ is bounded and that $(\mathbf{I} + \mathbf{C}\mathbf{P})$ has closed graph and domain equal to $\mathcal{D}_{\mathbf{P}}$. This gives that $\mathbf{P} = \mathbf{P}(\mathbf{I} + \mathbf{C}\mathbf{P})^{-1}(\mathbf{I} + \mathbf{C}\mathbf{P})$ has closed graph.

The following proposition presents a geometric characterization of stability of $[\mathbf{P}, \mathbf{C}]$.

PROPOSITION 2. *Let $\mathbf{P} \in \mathcal{P}_{\mathcal{U}, \mathcal{Y}}$ and $\mathbf{C} \in \mathcal{P}_{\mathcal{Y}, \mathcal{U}}$. Then the following are equivalent:*

- (a) $[\mathbf{P}, \mathbf{C}]$ is stable,
- (b) $\mathcal{G}_{\mathbf{P}}, \mathcal{G}'_{\mathbf{C}}$ are closed,

$$(1) \quad \mathcal{G}_{\mathbf{P}} \cap \mathcal{G}'_{\mathbf{C}} = \{0\}$$

and

$$(2) \quad \mathcal{G}_{\mathbf{P}} + \mathcal{G}'_{\mathbf{C}} = \mathcal{L}.$$

Proof. (a) \Rightarrow (b). If $[\mathbf{P}, \mathbf{C}]$ is stable then, by Proposition 1, $\mathcal{G}_{\mathbf{P}}, \mathcal{G}'_{\mathbf{C}}$ are closed. Moreover, both (1) and (2) are necessary for $\mathbf{F}_{\mathbf{P}, \mathbf{C}}$ to be a one-to-one mapping onto \mathcal{L} .

(b) \Rightarrow (a). First note that (1) and (2) are sufficient to guarantee the existence of a set-theoretic inverse for $\mathbf{F}_{\mathbf{P}, \mathbf{C}}$ defined on \mathcal{L} . We need to show that the inverse is also bounded. We first observe that the graph of $\mathbf{F}_{\mathbf{P}, \mathbf{C}}$ is closed. To see this note that

$$\begin{aligned} \mathcal{G}_{\mathbf{F}} &= \left\{ \begin{pmatrix} e_1 \\ -e_2 \\ e_1 - \mathbf{C}e_2 \\ \mathbf{P}e_1 - e_2 \end{pmatrix} : \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \in \mathcal{D}_{\mathbf{P}} \times \mathcal{D}_{\mathbf{C}} \right\} \\ &= \left\{ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} e_1 \\ \mathbf{P}e_1 \\ \mathbf{C}e_2 \\ e_2 \end{pmatrix} : \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \in \mathcal{D}_{\mathbf{P}} \times \mathcal{D}_{\mathbf{C}} \right\} \\ &\cong \left\{ \begin{pmatrix} e_1 \\ \mathbf{P}e_1 \\ \mathbf{C}e_2 \\ e_2 \end{pmatrix} : \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \in \mathcal{D}_{\mathbf{P}} \times \mathcal{D}_{\mathbf{C}} \right\} = \begin{pmatrix} \mathcal{G}_{\mathbf{P}} \\ \mathcal{G}'_{\mathbf{C}} \end{pmatrix}, \end{aligned}$$

where \cong denotes a Hilbert space isomorphism. Consequently, the graph of $\mathbf{H}_{\mathbf{P},\mathbf{C}}$ ($= \mathbf{F}_{\mathbf{P},\mathbf{C}}^{-1}$) is also closed. The result now follows from the closed graph theorem. \square

Similar geometric concepts for expressing stability have been employed in earlier studies, notably in the context of nonlinear control systems [18], [22], and in the recent works [3], [14], [16], [19].

3. Preliminaries on the gap metric. In light of Proposition 2 we will restrict our attention in the rest of the paper to linear systems that have closed graphs. We will identify \mathbf{P} (through its graph) and \mathbf{C} (through its inverse graph) with elements of

$$S_{\mathcal{L}} := \{\mathcal{K} : \mathcal{K} \text{ is a closed subspace of } \mathcal{L}\}.$$

For $\mathcal{K} \in S_{\mathcal{L}}$ denote by $\Pi_{\mathcal{K}}$ the orthogonal projection with range \mathcal{K} . The *gap* between $\mathcal{K}_1, \mathcal{K}_2 \in S_{\mathcal{L}}$ is the metric defined as

$$\delta(\mathcal{K}_1, \mathcal{K}_2) := \|\Pi_{\mathcal{K}_1} - \Pi_{\mathcal{K}_2}\|$$

(see [11] and [12]), and $S_{\mathcal{L}}$ is equipped with the natural topology induced by the gap metric. Thus the *gap* between two systems \mathbf{P}_i , $i = 1, 2$, is defined to be the gap between their respective graphs $\mathcal{G}_{\mathbf{P}_i}$, $i = 1, 2$ ([25]).

Let $\mathcal{B}(\mathcal{K}_1, \mathcal{K}_2)$ denote the space of bounded operators between two Hilbert spaces \mathcal{K}_1 and \mathcal{K}_2 . For $\mathbf{X} \in \mathcal{B}(\mathcal{K}_1, \mathcal{K}_2)$ define $\tau(\mathbf{X}) := \inf_{x \in \mathcal{K}_1, \|x\|=1} \|\mathbf{X}x\|$.

PROPOSITION 3. *Let $\mathcal{K}_0 \in S_{\mathcal{L}}$. The following are equivalent:*

- (a) $\mathcal{K} \in S_{\mathcal{L}}$ and $\delta(\mathcal{K}_0, \mathcal{K}) < 1$;
- (b) $\Pi_{\mathcal{K}_0}|_{\mathcal{K}}$ is invertible;
- (c) There exists an $\mathbf{X} \in \mathcal{B}(\mathcal{K}_0, \mathcal{K}_0^\perp)$ such that $\mathcal{K} = (\mathbf{I}_{\mathcal{K}_0} + \mathbf{X})\mathcal{K}_0$.

Furthermore, if $\mathcal{K} \in S_{\mathcal{L}}$ and $\delta(\mathcal{K}_0, \mathcal{K}) < 1$, then

$$\begin{aligned} (3) \quad \delta(\mathcal{K}_0, \mathcal{K}) &= \sqrt{1 - \tau^2(\Pi_{\mathcal{K}_0}|_{\mathcal{K}})} = \sqrt{1 - \tau^2(\Pi_{\mathcal{K}}|_{\mathcal{K}_0})} \\ (4) \quad &= \|\mathbf{X}(\mathbf{I} + \mathbf{X}^*\mathbf{X})^{-1/2}\| = \frac{\|\mathbf{X}\|}{\sqrt{1 + \|\mathbf{X}\|^2}}. \end{aligned}$$

Proof. The equivalence (a) \Leftrightarrow (b), as well as a proof of (3), can be found in [12, Lemma 15.1].

(b) \Rightarrow (c). Since $\delta(\mathcal{K}_0, \mathcal{K}) = \delta(\mathcal{K}, \mathcal{K}_0)$ both $\Pi_{\mathcal{K}_0}|_{\mathcal{K}}$ and $\Pi_{\mathcal{K}}|_{\mathcal{K}_0}$ are invertible. Hence

$$\begin{aligned} \mathcal{K} &= \Pi_{\mathcal{K}}|_{\mathcal{K}_0}\mathcal{K}_0 = \left(\Pi_{\mathcal{K}_0}\Pi_{\mathcal{K}} + \Pi_{\mathcal{K}_0^\perp}\Pi_{\mathcal{K}}\right)\mathcal{K}_0 \\ &= \left(\mathbf{I}_{\mathcal{K}_0} + \Pi_{\mathcal{K}_0^\perp}\Pi_{\mathcal{K}}(\Pi_{\mathcal{K}_0}|_{\mathcal{K}})^{-1}\right)\mathcal{K}_0. \end{aligned}$$

Thus, (c) holds for $\mathbf{X} = \Pi_{\mathcal{K}_0^\perp}(\Pi_{\mathcal{K}_0}|_{\mathcal{K}})^{-1}$.

(c) \Rightarrow (b). From the equality $k_0 = \Pi_{\mathcal{K}_0}|_{\mathcal{K}}(\mathbf{I} + \mathbf{X})k_0$, for $k_0 \in \mathcal{K}_0$, we see that $\Pi_{\mathcal{K}_0}|_{\mathcal{K}}$ is onto and from $\mathcal{K} = (\mathbf{I}_{\mathcal{K}_0} + \mathbf{X})\mathcal{K}_0$ that it is one-to-one.

We defer the derivation of (4) to Proposition 4 below, where we prove a slightly more general statement. \square

From the definition of the gap metric it follows that

$$\begin{aligned} \delta(\mathcal{K}_1, \mathcal{K}_2) &= \left\| \begin{pmatrix} \Pi_{\mathcal{K}_1} \\ \Pi_{\mathcal{K}_1^\perp} \end{pmatrix} (\Pi_{\mathcal{K}_1} - \Pi_{\mathcal{K}_2}) (\Pi_{\mathcal{K}_2^\perp}, \Pi_{\mathcal{K}_2}) \right\| \\ &= \left\| \begin{pmatrix} \Pi_{\mathcal{K}_1}\Pi_{\mathcal{K}_2^\perp} & \mathbf{0} \\ \mathbf{0} & -\Pi_{\mathcal{K}_1^\perp}\Pi_{\mathcal{K}_2} \end{pmatrix} \right\| \\ &= \max\{\|\Pi_{\mathcal{K}_1^\perp}\Pi_{\mathcal{K}_2}\|, \|\Pi_{\mathcal{K}_2^\perp}\Pi_{\mathcal{K}_1}\|\} \end{aligned}$$

(see [12]). Note that

$$\|\Pi_{\mathcal{K}_1^\perp} \Pi_{\mathcal{K}_2}\| = \sup_{x \in \mathcal{K}_2, \|x\|=1} \text{dist}(x, \mathcal{K}_1),$$

where $\text{dist}(x, \mathcal{K}_1) := \inf_{y \in \mathcal{K}_1} \|x - y\|$.

PROPOSITION 4. Let $\mathbf{X}_i \in \mathcal{B}(\mathcal{K}_0, \mathcal{K}_0^\perp)$, for $i = 1, 2$, and $\mathcal{K}_i = (\mathbf{I}_{\mathcal{K}_0} + \mathbf{X}_i)\mathcal{K}_0$. Then

$$\delta(\mathcal{K}_1, \mathcal{K}_2) = \max \left\{ \left\| (\mathbf{I} + \mathbf{X}_1 \mathbf{X}_1^*)^{-1/2} (\mathbf{X}_2 - \mathbf{X}_1) (\mathbf{I} + \mathbf{X}_2^* \mathbf{X}_2)^{-1/2} \right\|, \right. \\ \left. \left\| (\mathbf{I} + \mathbf{X}_2 \mathbf{X}_2^*)^{-1/2} (\mathbf{X}_2 - \mathbf{X}_1) (\mathbf{I} + \mathbf{X}_1^* \mathbf{X}_1)^{-1/2} \right\| \right\} \quad (5)$$

$$= \sqrt{1 - \rho^2}, \quad (6)$$

where

$$\rho = \min \left\{ \tau \left((\mathbf{I} + \mathbf{X}_1^* \mathbf{X}_1)^{-1/2} (\mathbf{I} + \mathbf{X}_1^* \mathbf{X}_2) (\mathbf{I} + \mathbf{X}_2^* \mathbf{X}_2)^{-1/2} \right), \right. \\ \left. \tau \left((\mathbf{I} + \mathbf{X}_1 \mathbf{X}_1^*)^{-1/2} (\mathbf{I} + \mathbf{X}_1 \mathbf{X}_2^*) (\mathbf{I} + \mathbf{X}_2 \mathbf{X}_2^*)^{-1/2} \right) \right\}.$$

Proof. We compute

$$\|\Pi_{\mathcal{K}_1^\perp} \Pi_{\mathcal{K}_2}\| = \left\| \begin{pmatrix} -\mathbf{X}_1^* \\ \mathbf{I} \end{pmatrix} (\mathbf{I} + \mathbf{X}_1 \mathbf{X}_1^*)^{-1} (-\mathbf{X}_1, \mathbf{I}) \begin{pmatrix} \mathbf{I} \\ \mathbf{X}_2 \end{pmatrix} (\mathbf{I} + \mathbf{X}_2^* \mathbf{X}_2)^{-1} (\mathbf{I}, \mathbf{X}_2^*) \right\| \\ = \|(\mathbf{I} + \mathbf{X}_1 \mathbf{X}_1^*)^{-1/2} (\mathbf{X}_2 - \mathbf{X}_1) (\mathbf{I} + \mathbf{X}_2^* \mathbf{X}_2)^{-1/2}\| \quad (7)$$

since

$$\begin{pmatrix} -\mathbf{X}_1^* \\ \mathbf{I} \end{pmatrix} (\mathbf{I} + \mathbf{X}_1 \mathbf{X}_1^*)^{-1/2}$$

is an isometry and $(\mathbf{I} + \mathbf{X}_2^* \mathbf{X}_2)^{-1/2} (\mathbf{I}, \mathbf{X}_2^*)$ is a co-isometry. By symmetry $\|\Pi_{\mathcal{K}_2^\perp} \Pi_{\mathcal{K}_1}\|$ is given by the dual expression. This completes the proof of (5).

To prove (6) consider the unitary operators

$$\mathbf{Y}_i := \begin{pmatrix} (\mathbf{I} + \mathbf{X}_i^* \mathbf{X}_i)^{-1/2} & -\mathbf{X}_i^* (\mathbf{I} + \mathbf{X}_i \mathbf{X}_i^*)^{-1/2} \\ \mathbf{X}_i (\mathbf{I} + \mathbf{X}_i^* \mathbf{X}_i)^{-1/2} & (\mathbf{I} + \mathbf{X}_i \mathbf{X}_i^*)^{-1/2} \end{pmatrix},$$

for $i = 1, 2$, and define

$$\mathbf{Y} := \mathbf{Y}_1^* \mathbf{Y}_2 = \begin{pmatrix} (\mathbf{Y})_{1,1} & (\mathbf{Y})_{1,2} \\ (\mathbf{Y})_{2,1} & (\mathbf{Y})_{2,2} \end{pmatrix}, \quad (8)$$

where $(\mathbf{Y})_{i,j}$ denotes the (i, j) -block entry of \mathbf{Y} . Since

$$\begin{pmatrix} (\mathbf{Y})_{1,1} \\ (\mathbf{Y})_{2,1} \end{pmatrix}$$

is an isometry, it follows that $\|(\mathbf{Y})_{2,1}\|^2 + \tau^2((\mathbf{Y})_{1,1}) = 1$. Using (7) it follows that

$$\|\Pi_{\mathcal{K}_1^\perp} \Pi_{\mathcal{K}_2}\| = \|(\mathbf{Y})_{2,1}\| \\ = \sqrt{1 - \tau^2((\mathbf{Y})_{1,1})},$$

where $(\mathbf{Y})_{1,1} = (\mathbf{I} + \mathbf{X}_1^* \mathbf{X}_1)^{-1/2} (\mathbf{I} + \mathbf{X}_1^* \mathbf{X}_2) (\mathbf{I} + \mathbf{X}_2^* \mathbf{X}_2)^{-1/2}$. Similarly, $\|\Pi_{\mathcal{K}_1^\perp} \Pi_{\mathcal{K}_2}\| = \sqrt{1 - \tau^2((\mathbf{Y})_{2,2})}$, where $(\mathbf{Y})_{2,2} = (\mathbf{I} + \mathbf{X}_1 \mathbf{X}_1^*)^{-1/2} (\mathbf{I} + \mathbf{X}_1 \mathbf{X}_2^*) (\mathbf{I} + \mathbf{X}_2 \mathbf{X}_2^*)^{-1/2}$. This completes the proof. \square

Consider two subspaces \mathcal{K}_0 and \mathcal{K}_1 at a distance $\delta(\mathcal{K}_0, \mathcal{K}_1) < 1$, where $\mathcal{K}_1 = (\mathbf{I} + \mathbf{X})\mathcal{K}$ and $\mathbf{X} \in \mathcal{B}(\mathcal{K}_0, \mathcal{K}_0^\perp)$. Define

$$(9) \quad \mathcal{K}_\lambda = (\mathbf{I} + \lambda \mathbf{X})\mathcal{K}_0$$

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

COROLLARY 1. *The family \mathcal{K}_λ , $\lambda \in [0, 1]$, defines a path, continuous in the gap metric, between \mathcal{K}_0 and \mathcal{K}_1 . Moreover, for $\lambda \in \mathbb{R}$,*

$$(10) \quad \delta(\mathcal{K}_0, \mathcal{K}_\lambda) = |\lambda| \|\mathbf{X}\| (1 + \lambda^2 \|\mathbf{X}\|^2)^{-1/2},$$

$$(11) \quad \delta(\mathcal{K}_1, \mathcal{K}_\lambda) = \|1 - \lambda\| \|\mathbf{X}^* \mathbf{X} (1 + \mathbf{X}^* \mathbf{X})^{-1} (1 + \lambda^2 \mathbf{X}^* \mathbf{X})^{-1}\|^{1/2}$$

$$(12) \quad \leq \frac{|1 - \lambda|}{|1 + \lambda|}.$$

Proof. To establish that the family \mathcal{K}_λ , $\lambda \in [0, 1]$, defines a path it suffices to prove (10)–(12). Equation (10) follows from Proposition 4 by identifying $\mathcal{K}_0, \mathcal{K}_1, \mathcal{K}_2$ with $\mathcal{K}_0, \mathcal{K}_0, \mathcal{K}_\lambda$, respectively, and $\mathbf{X}_1, \mathbf{X}_2$ with $\mathbf{0}, \lambda \mathbf{X}$. If, in Proposition 4, we identify $\mathcal{K}_0, \mathcal{K}_1, \mathcal{K}_2$ with $\mathcal{K}_0, \mathcal{K}_1, \mathcal{K}_\lambda$ and $\mathbf{X}_1, \mathbf{X}_2$ with $\mathbf{X}, \lambda \mathbf{X}$, then we obtain

$$(13) \quad \delta(\mathcal{K}_1, \mathcal{K}_\lambda) = \max \left\{ \|1 - \lambda\| \|(1 + \mathbf{X} \mathbf{X}^*)^{-1/2} \mathbf{X} (1 + \lambda^2 \mathbf{X}^* \mathbf{X})^{-1/2}\|, \right. \\ \left. \|1 - \lambda\| \|(1 + \mathbf{X} \mathbf{X}^*)^{-1/2} \mathbf{X} (1 + \lambda^2 \mathbf{X}^* \mathbf{X})^{-1/2}\| \right\}.$$

Both expressions are equal and can be seen to equal the right-hand side of (11). Since $\mathbf{X}^* \mathbf{X}$ is a positive, bounded, selfadjoint operator, by invoking the spectral mapping theorem, it follows that

$$\delta(\mathcal{K}_1, \mathcal{K}_\lambda) = \sup \left\{ \frac{|1 - \lambda| \sqrt{x}}{\sqrt{(1 + x)(1 + \lambda^2 x)}} : x \in \text{Spectrum}(\mathbf{X}^* \mathbf{X}) \right\}.$$

The supremum of the function in the interval $[0, \infty)$ occurs at $x = 1/\lambda$ and equals $|1 - \lambda|/|1 + \lambda|$. (Note that $\text{Spectrum}(\mathbf{X}^* \mathbf{X}) \subset [0, \infty)$.) \square

COROLLARY 2. *Let $\mathcal{K}_0, \mathcal{K}_1, \mathcal{K}_2 \in \mathcal{S}_{\mathcal{L}}$ be such that $\delta^2(\mathcal{K}_0, \mathcal{K}_1) + \delta^2(\mathcal{K}_0, \mathcal{K}_2) < 1$. Then*

$$(14) \quad \delta(\mathcal{K}_1, \mathcal{K}_2) \leq \delta(\mathcal{K}_0, \mathcal{K}_1) \sqrt{1 - \delta^2(\mathcal{K}_0, \mathcal{K}_2)} + \delta(\mathcal{K}_0, \mathcal{K}_2) \sqrt{1 - \delta^2(\mathcal{K}_0, \mathcal{K}_1)}.$$

Proof. Since $\delta(\mathcal{K}_0, \mathcal{K}_1) < 1$ and $\delta(\mathcal{K}_0, \mathcal{K}_2) < 1$, by Proposition 3 there exist bounded operators $\mathbf{X}_i : \mathcal{K}_0 \rightarrow \mathcal{K}_0^\perp$ such that $\mathcal{K}_i = (\mathbf{I}_{\mathcal{K}_0} + \mathbf{X}_i)\mathcal{K}_0$, for $i = 1, 2$. We observe that

$$\delta^2(\mathcal{K}_0, \mathcal{K}_1) + \delta^2(\mathcal{K}_0, \mathcal{K}_2) < 1 \Rightarrow \frac{\|\mathbf{X}_1\|^2}{1 + \|\mathbf{X}_1\|^2} + \frac{\|\mathbf{X}_2\|^2}{1 + \|\mathbf{X}_2\|^2} < 1 \\ \Rightarrow \|\mathbf{X}_1\| \|\mathbf{X}_2\| < 1.$$

Since $\|\mathbf{X}_1^*\mathbf{X}_2\| \leq \|\mathbf{X}_1\|\|\mathbf{X}_2\| < 1$, it follows that $\tau(\mathbf{I} + \mathbf{X}_1^*\mathbf{X}_2) = \tau(\mathbf{I} + \mathbf{X}_2^*\mathbf{X}_1) > 0$. Consequently $\delta(\mathcal{K}_1, \mathcal{K}_2) = \sqrt{1 - \rho^2} < 1$, where

$$\begin{aligned} \rho &= \tau \left((\mathbf{I} + \mathbf{X}_1^*\mathbf{X}_1)^{-1/2} (\mathbf{I} + \mathbf{X}_1^*\mathbf{X}_2) (\mathbf{I} + \mathbf{X}_2^*\mathbf{X}_2)^{-1/2} \right) \\ &\geq \tau \left((\mathbf{I} + \mathbf{X}_1^*\mathbf{X}_1)^{-1/2} \right) \tau(\mathbf{I} + \mathbf{X}_1^*\mathbf{X}_2) \tau \left((\mathbf{I} + \mathbf{X}_2^*\mathbf{X}_2)^{-1/2} \right) \\ &\geq \frac{1 - \|\mathbf{X}_1\|\|\mathbf{X}_2\|}{\sqrt{1 + \|\mathbf{X}_1\|^2} \sqrt{1 + \|\mathbf{X}_2\|^2}}. \end{aligned}$$

Thus,

$$\begin{aligned} \delta(\mathcal{K}_1, \mathcal{K}_2) &\leq \sqrt{1 - \frac{(1 - \|\mathbf{X}_1\|\|\mathbf{X}_2\|)^2}{(1 + \|\mathbf{X}_1\|^2)(1 + \|\mathbf{X}_2\|^2)}} \\ &= \frac{\|\mathbf{X}_1\|}{\sqrt{1 + \|\mathbf{X}_1\|^2}} \frac{1}{\sqrt{1 + \|\mathbf{X}_2\|^2}} + \frac{\|\mathbf{X}_2\|}{\sqrt{1 + \|\mathbf{X}_2\|^2}} \frac{1}{\sqrt{1 + \|\mathbf{X}_1\|^2}}. \end{aligned}$$

This completes the proof. \square

The arcsine of the gap metric can be thought of as the *maximal angle* between two subspaces, denoted by $\theta_{\max}(\mathcal{K}_1, \mathcal{K}_2) := \arcsin \delta(\mathcal{K}_1, \mathcal{K}_2)$. Corollary 2 is given in [16]. It was also observed in [16] that (14) can be rewritten in the form

$$(15) \quad \theta_{\max}(\mathcal{K}_1, \mathcal{K}_2) \leq \theta_{\max}(\mathcal{K}_1, \mathcal{K}) + \theta_{\max}(\mathcal{K}, \mathcal{K}_2)$$

so that θ_{\max} defines a metric in $S_{\mathcal{L}}$.

Given any subspace $\mathcal{K}_0 \in S_{\mathcal{L}}$ and a positive number b let

$$\text{Ball}(\mathcal{K}_0, b) := \{\mathcal{K} : \delta(\mathcal{K}_0, \mathcal{K}) < b\}$$

denote the gap-ball about \mathcal{K}_0 of radius b , and let $\overline{\text{Ball}}(\mathcal{K}_0, b)$ denote the closure of $\text{Ball}(\mathcal{K}_0, b)$.

PROPOSITION 5. *If $b < 1$, then $\overline{\text{Ball}}(\mathcal{K}_0, b) = \{\mathcal{K} : \delta(\mathcal{K}_0, \mathcal{K}) \leq b\}$.*

Proof. Let \mathcal{K}_1 be such that $\delta(\mathcal{K}_0, \mathcal{K}_1) = b$ and consider \mathcal{K}_λ , $\lambda \in \mathbb{R}$, constructed as in Corollary 1. It follows that

$$\{\mathcal{K}_\lambda : \lambda \in [0, 1)\} \subset \text{Ball}(\mathcal{K}_0, b).$$

Thus, any neighbourhood of \mathcal{K}_1 contains a subspace \mathcal{K}_λ for some $\lambda \in [0, 1)$. Hence, $\mathcal{K}_1 \in \overline{\text{Ball}}(\mathcal{K}_0, b)$. Conversely, take \mathcal{K}_1 such that for all $\epsilon > 0$, $\text{Ball}(\mathcal{K}_0, b) \cap \text{Ball}(\mathcal{K}_1, \epsilon) \neq \emptyset$. Then take $\mathcal{K} \in \text{Ball}(\mathcal{K}_0, b) \cap \text{Ball}(\mathcal{K}_1, \epsilon)$. Using the triangular inequality it follows that $\delta(\mathcal{K}_0, \mathcal{K}_1) \leq \delta(\mathcal{K}_0, \mathcal{K}) + \delta(\mathcal{K}, \mathcal{K}_1) < b + \epsilon$. Since this is valid for any $\epsilon > 0$, $\delta(\mathcal{K}_0, \mathcal{K}_1) \leq b$. \square

It should be noted that when $b = 1$, then $\overline{\text{Ball}}(\mathcal{K}_0, b) \neq \{\mathcal{K} : \delta(\mathcal{K}_0, \mathcal{K}) \leq b\}$.

THEOREM 1. *Let $b, \zeta \in \mathbb{R}$, with $0 < b < 1$ and $0 < \zeta < 1$. There exists $\epsilon > 0$ such that*

$$(16) \quad \text{Ball}(\mathcal{K}_0, b + \epsilon) \subseteq \bigcup_{\mathcal{K} \in \overline{\text{Ball}}(\mathcal{K}_0, b)} \text{Ball}(\mathcal{K}, \zeta).$$

Proof. Consider any \mathcal{K}_a such that $1 > \delta(\mathcal{K}_0, \mathcal{K}_a) = a > b$. We will construct a \mathcal{K}_b with $\delta(\mathcal{K}_0, \mathcal{K}_b) = b$ and $\delta(\mathcal{K}_b, \mathcal{K}_a) = a\sqrt{1 - b^2} - b\sqrt{1 - a^2}$. This establishes (16) for any ϵ such that $(b + \epsilon)\sqrt{1 - b^2} - b\sqrt{1 - (b + \epsilon)^2} < \zeta$.

Write $\mathcal{K}_a = (\mathbf{I} + \mathbf{X})\mathcal{K}_0$ with $\mathbf{X} \in \mathcal{B}(\mathcal{K}_0, \mathcal{K}_0^\perp)$. Let $\mathbf{X} = \mathbf{U}\mathbf{R}$ be a polar decomposition for \mathbf{X} (i.e., with $\mathbf{R} = (\mathbf{X}^*\mathbf{X})^{1/2}$ a positive selfadjoint operator and $\mathbf{U} : \text{range}\mathbf{R} \rightarrow \text{range}\mathbf{X}$ a partial isometry), denote by \mathbf{E}_λ the spectral family of projections corresponding to \mathbf{R} , define

$$\mathbf{\Lambda} := \int_{0-}^{\infty} g(\lambda) d\mathbf{E}_\lambda,$$

where

$$g(\lambda) = \begin{cases} 1 & \text{if } 0 \leq \lambda < \frac{b}{\sqrt{1-b^2}} \\ \frac{1}{\lambda} \frac{b}{\sqrt{1-b^2}} & \text{if } \frac{b}{\sqrt{1-b^2}} \leq \lambda \end{cases},$$

and define

$$(17) \quad \mathcal{K}_b := (\mathbf{I} + \mathbf{X}\mathbf{\Lambda})\mathcal{K}_0.$$

In the rest of the proof we verify that $\delta(\mathcal{K}_0, \mathcal{K}_b) = b$ and $\delta(\mathcal{K}_b, \mathcal{K}_a) = a\sqrt{1-b^2} - b\sqrt{1-a^2}$.

From Proposition 3, $\|\mathbf{X}\| = a/\sqrt{1-a^2}$. Also,

$$\begin{aligned} \|\mathbf{X}\mathbf{\Lambda}\| &= \|\mathbf{R}\mathbf{\Lambda}\| \\ &= \left\| \int_{0-}^{\infty} \lambda g(\lambda) d\mathbf{E}_\lambda \right\| \\ &= \sup\{\lambda g(\lambda) : \lambda \in \text{Spectrum}(\mathbf{R})\} \\ &= \lambda g(\lambda) \Big|_{\frac{a}{\sqrt{1-a^2}}} \\ &= \frac{b}{\sqrt{1-b^2}} \end{aligned}$$

since $\lambda g(\lambda)$ is nondecreasing on $[0, \infty)$, $\text{Spectrum}(\mathbf{R}) \subseteq [0, \|\mathbf{R}\|]$, and $(b/\sqrt{1-b^2}) < (a/\sqrt{1-a^2}) = \|\mathbf{R}\|$. From Proposition 3, we conclude that

$$\delta(\mathcal{K}_0, \mathcal{K}_b) = \frac{\|\mathbf{X}\mathbf{\Lambda}\|}{\sqrt{1 + \|\mathbf{X}\mathbf{\Lambda}\|^2}} = b.$$

From Proposition 4, we have that

$$\delta(\mathcal{K}_a, \mathcal{K}_b) = \max\{\delta_1, \delta_2\},$$

where the two expressions δ_1, δ_2 are computed below. First,

$$\begin{aligned} \delta_1 &:= \|(\mathbf{I} + \mathbf{X}\mathbf{X}^*)^{-1/2}(\mathbf{X}\mathbf{\Lambda} - \mathbf{X})(\mathbf{I} + \mathbf{\Lambda}\mathbf{X}^*\mathbf{X}\mathbf{\Lambda})^{-1/2}\| \\ &= \|\mathbf{U}\mathbf{R}(\mathbf{I} + \mathbf{R}^2)^{-1/2}(\mathbf{\Lambda} - \mathbf{I})(\mathbf{I} + \mathbf{\Lambda}\mathbf{R}^2\mathbf{\Lambda})^{-1/2}\| \\ &= \|\mathbf{R}(\mathbf{I} + \mathbf{R}^2)^{-1/2}(\mathbf{\Lambda} - \mathbf{I})(\mathbf{I} + \mathbf{\Lambda}\mathbf{R}^2\mathbf{\Lambda})^{-1/2}\| \\ (18) \quad &= \sup\left\{ \frac{\lambda|g(\lambda) - 1|}{\sqrt{1 + \lambda^2}\sqrt{1 + \lambda^2g(\lambda)^2}} : \lambda \in \text{Spectrum}(\mathbf{R}) \right\} \end{aligned}$$

$$\begin{aligned} (19) \quad &= \frac{\lambda|g(\lambda) - 1|}{\sqrt{1 + \lambda^2}\sqrt{1 + \lambda^2g(\lambda)^2}} \text{ evaluated at } \lambda = \frac{a}{\sqrt{1-a^2}} \\ &= a\sqrt{1-b^2} - b\sqrt{1-a^2}. \end{aligned}$$

The step (18) \Rightarrow (19) follows because $\lambda|g(\lambda) - 1|/\sqrt{1+\lambda^2}\sqrt{1+\lambda^2g(\lambda)^2}$ is monotonically nondecreasing in $\text{Spectrum}(\mathbf{R}) \subseteq [0, a/\sqrt{1-a^2}]$, while $\|\mathbf{R}\| = a/\sqrt{1-a^2}$. Next,

$$\begin{aligned} \delta_2 &:= \|(\mathbf{I} + \mathbf{X}\mathbf{\Lambda}^2\mathbf{X}^*)^{-1/2}(\mathbf{X}\mathbf{\Lambda} - \mathbf{X})(\mathbf{I} + \mathbf{X}^*\mathbf{X})^{-1/2}\| \\ (20) \quad &= \|(\mathbf{I} + \mathbf{U}\mathbf{R}\mathbf{\Lambda}^2\mathbf{R}\mathbf{U}^*)^{-1/2}\mathbf{U}\mathbf{R}(\mathbf{\Lambda} - \mathbf{I})(\mathbf{I} + \mathbf{R}^2)^{-1/2}\| \\ (21) \quad &= \|\mathbf{U}(\mathbf{I} + \mathbf{R}\mathbf{\Lambda}^2\mathbf{R})^{-1/2}\mathbf{R}(\mathbf{\Lambda} - \mathbf{I})(\mathbf{I} + \mathbf{R}^2)^{-1/2}\| \\ &= \delta_1 \end{aligned}$$

since \mathbf{R} and $\mathbf{\Lambda}$ commute. The step (20) \Rightarrow (21) is based on the fact that $\mathbf{R}\mathbf{U}^*\mathbf{U} = \mathbf{R}$ and $\mathbf{U}^*\mathbf{U}\mathbf{R} = \mathbf{R}$. \square

It is interesting to note that for arbitrary $\mathcal{K}_0, \mathcal{K}_a \in S_{\mathcal{L}}$ with $0 < b < a = \delta(\mathcal{K}_0, \mathcal{K}_a) < 1$ and \mathcal{K}_b as in (17), we have $\delta(\mathcal{K}_0, \mathcal{K}_b) = b$ and $\theta_{\max}(\mathcal{K}_0, \mathcal{K}_b) + \theta_{\max}(\mathcal{K}_b, \mathcal{K}_a) = \theta_{\max}(\mathcal{K}_0, \mathcal{K}_a)$.

4. Graphability. Let $\mathcal{K}, \mathcal{W} \in S_{\mathcal{L}}$. We say that \mathcal{K} is a *graph with respect to* \mathcal{W} if $\mathcal{K} \cap \mathcal{W}^\perp = \{0\}$. For any such \mathcal{K} we can define a linear operator \mathbf{K} by the relation $\mathbf{K}(\Pi_{\mathcal{W}}k) = \Pi_{\mathcal{W}^\perp}k$ for all $k \in \mathcal{K}$. For convenience we will identify each $u \in \mathcal{U}$ with $\begin{pmatrix} u \\ 0 \end{pmatrix} \in \mathcal{L}$ and, similarly, every $y \in \mathcal{Y}$ with $\begin{pmatrix} 0 \\ y \end{pmatrix} \in \mathcal{L}$, i.e., \mathcal{U} is identified with the subspace $\mathcal{U} \oplus \{0\}$ of \mathcal{L} and \mathcal{Y} with $\{0\} \oplus \mathcal{Y}$. We say \mathcal{K} is a *graph* if \mathcal{K} is a graph with respect to \mathcal{U} . Similarly, we say \mathcal{K} is an *inverse graph* if \mathcal{K} is a graph with respect to \mathcal{Y} . Denote

$$\text{Graph}_{\mathcal{W}} := \{\mathcal{K} \in S_{\mathcal{L}} : \mathcal{K} \cap \mathcal{W}^\perp = \{0\}\}.$$

For $\mathcal{K}, \mathcal{W} \in S_{\mathcal{L}}$ let $\mathbf{X}_{\mathcal{K}} := \Pi_{\mathcal{W}}|_{\mathcal{K}}$ and define S_{npi} to be the *complement* in $S_{\mathcal{L}}$ of the set

$$S_{\text{pi}} := \{\mathcal{K} \in S_{\mathcal{L}} : \mathbf{X}_{\mathcal{K}} \text{ is semi-Fredholm and } \text{ind } \mathbf{X}_{\mathcal{K}} > 0\}.$$

An operator \mathbf{X} is said to be *semi-Fredholm* if its range is closed and if at least one of $\dim \ker \mathbf{X}$, $\dim \ker \mathbf{X}^*$ is finite. In this case the *Fredholm index* is defined as $\text{ind } \mathbf{X} := \dim \ker \mathbf{X} - \dim \ker \mathbf{X}^*$.

LEMMA 1. S_{npi} is closed in $S_{\mathcal{L}}$.

Proof. Let $\mathcal{K} \in S_{\text{pi}}$ and let $\mathcal{K}_i \in S_{\mathcal{L}}$ satisfy $\delta(\mathcal{K}, \mathcal{K}_i) \rightarrow 0$ for $i = 1, 2, \dots$. We have

$$\begin{aligned} \|\mathbf{X}_{\mathcal{K}}^* - \Pi_{\mathcal{K}}\mathbf{X}_{\mathcal{K}_i}^*\| &= \|(\Pi_{\mathcal{K}} - \Pi_{\mathcal{K}}\Pi_{\mathcal{K}_i})|_{\mathcal{W}}\| \\ &\leq \|\Pi_{\mathcal{K}} - \Pi_{\mathcal{K}_i}\| = \delta(\mathcal{K}, \mathcal{K}_i) \rightarrow 0. \end{aligned}$$

Since the set of semi-Fredholm operators from \mathcal{W} to \mathcal{K} with a given index is open in the space of all bounded operators from \mathcal{W} to \mathcal{K} (see [11, Thm. 5.17]), it follows that there exists an N such that for $i \geq N$, $\Pi_{\mathcal{K}}\mathbf{X}_{\mathcal{K}_i}^*$ is semi-Fredholm and

$$\text{ind } \Pi_{\mathcal{K}}\mathbf{X}_{\mathcal{K}_i}^* = \text{ind } \mathbf{X}_{\mathcal{K}}^* = -\text{ind } \mathbf{X}_{\mathcal{K}}.$$

On the other hand, if $\delta(\mathcal{K}, \mathcal{K}_i) < 1$ then $\mathbf{Y} := \Pi_{\mathcal{K}}|_{\mathcal{K}_i}$ is an invertible operator from \mathcal{K}_i to \mathcal{K} and therefore $\mathbf{X}_{\mathcal{K}_i}^* = \mathbf{Y}^{-1}\Pi_{\mathcal{K}}\mathbf{X}_{\mathcal{K}_i}^*$ is semi-Fredholm and $\text{ind } \mathbf{X}_{\mathcal{K}_i}^* = \text{ind } \Pi_{\mathcal{K}}\mathbf{X}_{\mathcal{K}_i}^*$ for $i \geq N$. It follows that for i large enough $\mathbf{X}_{\mathcal{K}_i}^*$ is also semi-Fredholm and

$$\text{ind } \mathbf{X}_{\mathcal{K}_i} = -\text{ind } \mathbf{X}_{\mathcal{K}_i}^* = \text{ind } \mathbf{X}_{\mathcal{K}} > 0.$$

Thus S_{pi} is open and hence S_{npi} is closed in $S_{\mathcal{L}}$. \square

The following result characterizes the closure of $\text{Graph}_{\mathcal{W}}$ in $S_{\mathcal{L}}$. Any $\mathcal{K} \in \overline{\text{Graph}_{\mathcal{W}}}$ is said to be *graphable* with respect to \mathcal{W} .

THEOREM 2. $\overline{\text{Graph}}_{\mathcal{W}} = S_{\text{npi}}$.

Before we proceed with the proof of the theorem we provide a characterization of S_{npi} . By \mathbf{E}_λ (respectively, $\mathbf{E}_{*,\lambda}$), $\lambda \in \mathbb{R}$, we denote the spectral family associated with $\mathbf{X}_K^* \mathbf{X}_K$ (respectively, $\mathbf{X}_K \mathbf{X}_K^*$). Then

$$\mathbf{X}_K^* \mathbf{X}_K = \int_{0-}^{\infty} \lambda d\mathbf{E}_\lambda, \text{ and } \mathbf{X}_K \mathbf{X}_K^* = \int_{0-}^{\infty} \lambda d\mathbf{E}_{*,\lambda}$$

and the projections are chosen (strongly) continuous from the right (see [17]).

LEMMA 2. $S_{\text{npi}} = \{K : \dim \mathbf{E}_\lambda K \leq \dim \mathbf{E}_{*,\lambda} \mathcal{W}, \text{ for all } \lambda > 0 \text{ and } \lambda \text{ small enough}\}$.

Proof. (Inclusion \supset). It is a standard fact that $\mathbf{E}_0 K = \ker \mathbf{X}_K$, $\mathbf{E}_{*,0} \mathcal{W} = \ker \mathbf{X}_K^*$. If $K \in S_{\text{pi}}$ then we must have $\dim \mathbf{E}_0 K > \dim \mathbf{E}_{*,0} \mathcal{W}$. Since \mathbf{X}_K is semi-Fredholm then $\dim \mathbf{E}_\lambda K \rightarrow \dim \mathbf{E}_0 K$ and $\dim \mathbf{E}_{*,\lambda} \mathcal{W} \rightarrow \dim \mathbf{E}_{*,0} \mathcal{W}$ as $\lambda \rightarrow 0$. Hence, $\dim \mathbf{E}_\lambda K > \dim \mathbf{E}_{*,\lambda} \mathcal{W}$ for sufficiently small $\lambda > 0$.

(Inclusion \subset). Assume $K \in S_{\text{npi}}$ and that there exists a sequence $\lambda_i > 0$, $i = 1, 2, \dots$, $\lambda_i \rightarrow 0$, for which $\dim \mathbf{E}_{\lambda_i} K > \dim \mathbf{E}_{*,\lambda_i} \mathcal{W}$. Since $\dim \mathbf{E}_\lambda K$, $\dim \mathbf{E}_{*,\lambda} \mathcal{W}$ are nondecreasing in λ , $\dim \mathbf{E}_{*,\lambda_i} \mathcal{W}$ must be constant and finite—say equal to d —for i large enough. Thus 0 is isolated in the spectrum of $\mathbf{X}_K \mathbf{X}_K^*$ and $\dim \ker \mathbf{X}_K^* = \dim \ker \mathbf{X}_K \mathbf{X}_K^* = d$. It follows that \mathbf{X}_K is semi-Fredholm and that for λ_i small enough $\text{ind } \mathbf{X}_K = \dim \ker \mathbf{X}_K - \dim \ker \mathbf{X}_K^* = \dim \mathbf{E}_{\lambda_i} K - d > 0$. Thus $K \in S_{\text{pi}}$, which is a contradiction. \square

Proof of Theorem 2. We first show that $\overline{\text{Graph}}_{\mathcal{W}} \subseteq S_{\text{npi}}$. Let $K \notin S_{\text{npi}}$. Then

$$\begin{aligned} \text{ind } \mathbf{X}_K > 0 &\Rightarrow \dim \ker \mathbf{X}_K > 0 \\ &\Rightarrow K \cap \mathcal{W}^\perp \neq \{0\} \\ &\Rightarrow K \notin \text{Graph}_{\mathcal{W}}. \end{aligned}$$

Thus $\text{Graph}_{\mathcal{W}} \subseteq S_{\text{npi}}$ and, since S_{npi} is closed in $S_{\mathcal{L}}$ in the topology induced by the gap metric by Lemma 1, this implies that $K \notin \overline{\text{Graph}}_{\mathcal{W}}$ and completes the proof of the first part.

We now show that $\overline{\text{Graph}}_{\mathcal{W}} \supseteq S_{\text{npi}}$. Let $K \in S_{\text{npi}}$. For $\epsilon > 0$ small enough $\dim \mathbf{E}_\epsilon K \leq \dim \mathbf{E}_{*,\epsilon} \mathcal{W}$ and therefore there exists an isometry $\mathbf{V}_\epsilon : \mathbf{E}_\epsilon K \rightarrow \mathbf{E}_{*,\epsilon} \mathcal{W}$. Set $\mathcal{H}_\epsilon := \mathbf{E}_\epsilon K$, $\mathcal{K}_\epsilon := (K \ominus \mathcal{H}_\epsilon) + (\mathbf{I} + \sqrt{2\epsilon} \mathbf{V}_\epsilon) \mathcal{H}_\epsilon$ and note that

$$\begin{aligned} \delta(K, \mathcal{K}_\epsilon) &= \delta \left((K \ominus \mathcal{H}_\epsilon) \oplus \mathcal{H}_\epsilon, (K \ominus \mathcal{H}_\epsilon) + (\mathbf{I} + \sqrt{2\epsilon} \mathbf{V}_\epsilon) \mathcal{H}_\epsilon \right) \\ &= \delta \left(\mathcal{H}_\epsilon, (\mathbf{I} + \sqrt{2\epsilon} \Pi_{(K \ominus \mathcal{H}_\epsilon)^\perp} \mathbf{V}_\epsilon) \mathcal{H}_\epsilon \right). \end{aligned}$$

Therefore, $\lim_{\epsilon \rightarrow 0} \delta(K, \mathcal{K}_\epsilon) = 0$ by Proposition 3. We will now show that $\mathcal{K}_\epsilon \in \text{Graph}_{\mathcal{W}}$. Let $y \in \mathcal{K}_\epsilon \cap \mathcal{W}^\perp$. Then $y = k + h + \sqrt{2\epsilon} \mathbf{V}_\epsilon h$ with $k \in K \ominus \mathcal{H}_\epsilon$, $h \in \mathcal{H}_\epsilon$. It follows that

$$\begin{aligned} 0 &= \Pi_{\mathcal{W}} y = \Pi_{\mathcal{W}} k + \Pi_{\mathcal{W}} h + \sqrt{2\epsilon} \mathbf{V}_\epsilon h \\ &= \mathbf{X}_K k + \mathbf{X}_K h + \sqrt{2\epsilon} \mathbf{V}_\epsilon h, \end{aligned}$$

and thus

$$-\mathbf{X}_K^* \mathbf{X}_K k = \mathbf{X}_K^* \mathbf{X}_K h + \sqrt{2\epsilon} \mathbf{X}_K^* \mathbf{V}_\epsilon h.$$

Since $\mathbf{E}_\epsilon \mathbf{X}_K^* \mathbf{X}_K = \mathbf{X}_K^* \mathbf{X}_K \mathbf{E}_\epsilon$, $\mathbf{E}_\epsilon \mathbf{X}_K^* = \mathbf{X}_K^* \mathbf{E}_{*,\epsilon}$ and $\mathbf{E}_{*,\epsilon} \mathbf{V}_\epsilon = \mathbf{V}_\epsilon$, we see that

$$0 = -\mathbf{E}_\epsilon \mathbf{X}_K^* \mathbf{X}_K k$$

$$\begin{aligned}
&= \mathbf{E}_\epsilon(\mathbf{X}_\mathcal{K}^* \mathbf{X}_\mathcal{K} h + \sqrt{2\epsilon} \mathbf{X}_\mathcal{K}^* \mathbf{V}_\epsilon h) \\
&= \mathbf{X}_\mathcal{K}^* \mathbf{X}_\mathcal{K} h + \sqrt{2\epsilon} \mathbf{X}_\mathcal{K}^* \mathbf{V}_\epsilon h \\
&= -\mathbf{X}_\mathcal{K}^* \mathbf{X}_\mathcal{K} k.
\end{aligned}$$

Since $\|\mathbf{X}_\mathcal{K}^* \mathbf{X}_\mathcal{K} k\| \geq \epsilon \|k\|$, it follows that $k = 0$. Thus $y = h + \sqrt{2\epsilon} \mathbf{V}_\epsilon h$ and therefore

$$2\epsilon \|h\|^2 = \|\sqrt{2\epsilon} \mathbf{V}_\epsilon h\|^2 = \|\mathbf{X}_\mathcal{K} h\|^2 \leq \epsilon \|h\|^2$$

because $h \in \mathcal{H}_\epsilon$. This implies that $h = 0$, that is $y = 0$. \square

Remark. It is easy to see that the complement of $\overline{\text{Graph}}_\mathcal{W}$ can be characterized in the following way: $\mathcal{K} \notin \overline{\text{Graph}}_\mathcal{W} \Leftrightarrow$ there exists $\epsilon > 0$ such that $\mathcal{K}' \cap \mathcal{W}^\perp \neq \{0\}$ for all $\mathcal{K}' \in \text{Ball}(\mathcal{K}, \epsilon)$. From Theorem 2 the complement of $\overline{\text{Graph}}_\mathcal{W}$ is the set S_{pi} . A geometric characterization of S_{pi} is as follows:

$$\text{S}_{\text{pi}} = \{\mathcal{K} \in \mathcal{S}_\mathcal{L} : \mathcal{K} + \mathcal{W}^\perp \text{ is closed, } \dim(\mathcal{K} \cap \mathcal{W}^\perp) > \dim(\mathcal{L} \ominus (\mathcal{K} + \mathcal{W}^\perp))\}.$$

To see the equality, note that $\mathcal{K} + \mathcal{W}^\perp = \Pi_\mathcal{W} \mathcal{K} + \mathcal{W}^\perp$. Hence $\mathcal{K} + \mathcal{W}^\perp$ is closed $\Leftrightarrow \Pi_\mathcal{W} \mathcal{K}$ is closed. Also $\ker(\Pi_\mathcal{W}|_\mathcal{K}) = \mathcal{K} \cap \mathcal{W}^\perp$ and $\ker(\Pi_\mathcal{K}|_\mathcal{W}) = \mathcal{W} \cap \mathcal{K}^\perp = \mathcal{L} \ominus (\mathcal{K} + \mathcal{W}^\perp)$.

5. Robust stabilization. Consider the feedback interconnection $[\mathbf{P}, \mathbf{C}]$ and let $\mathcal{M} = \mathcal{G}_\mathbf{P}$, $\mathcal{N} = \mathcal{G}'_\mathbf{C} \in \mathcal{S}_\mathcal{L}$. Define $\mathbf{A}_{\mathcal{M}, \mathcal{N}} := \Pi_{\mathcal{N}^\perp}|_\mathcal{M}$. The following is a standard result in operator theory.

PROPOSITION 6. *Let $\mathcal{M}, \mathcal{N} \in \mathcal{S}_\mathcal{L}$. The following are equivalent:*

- (a) $\mathcal{M} \cap \mathcal{N} = \{0\}$ and $\mathcal{M} + \mathcal{N} = \mathcal{L}$;
- (b) $\mathbf{A}_{\mathcal{M}, \mathcal{N}}$ is invertible.

Proof. (a) \Rightarrow (b). Note that

$$\begin{aligned}
\Pi_{\mathcal{N}^\perp} \mathcal{M} &= \Pi_{\mathcal{N}^\perp} (\mathcal{M} + \mathcal{N}) \\
&= \Pi_{\mathcal{N}^\perp} \mathcal{L} \\
&= \mathcal{N}^\perp.
\end{aligned}$$

Thus $\mathbf{A}_{\mathcal{M}, \mathcal{N}}$ maps \mathcal{M} onto \mathcal{N}^\perp . Since $\mathcal{M} \cap \mathcal{N} = \{0\}$ then $\mathbf{A}_{\mathcal{M}, \mathcal{N}}$ is one-to-one. Thus $\mathbf{A}_{\mathcal{M}, \mathcal{N}}$ is invertible (see [9, Prob. 52]).

(b) \Rightarrow (a). For any $x \in \mathcal{M} \cap \mathcal{N}$ it clearly holds that $\mathbf{A}_{\mathcal{M}, \mathcal{N}} x = 0$. Since $\mathbf{A}_{\mathcal{M}, \mathcal{N}}$ is invertible, then $\mathcal{M} \cap \mathcal{N} = \{0\}$. Also, for any $x \in \mathcal{L}$ we can write

$$(22) \quad x = \mathbf{A}_{\mathcal{M}, \mathcal{N}}^{-1} \Pi_{\mathcal{N}^\perp} x + (\mathbf{I}_\mathcal{L} - \mathbf{A}_{\mathcal{M}, \mathcal{N}}^{-1} \Pi_{\mathcal{N}^\perp}) x =: m + n.$$

Clearly $m = \mathbf{A}_{\mathcal{M}, \mathcal{N}}^{-1} \Pi_{\mathcal{N}^\perp} x \in \mathcal{M}$. We claim that $n \in \mathcal{N}$. To see this note that $\Pi_{\mathcal{N}^\perp} n = \Pi_{\mathcal{N}^\perp} x - \Pi_{\mathcal{N}^\perp} \mathbf{A}_{\mathcal{M}, \mathcal{N}}^{-1} \Pi_{\mathcal{N}^\perp} x = 0$. Thus, $\mathcal{L} = \mathcal{M} + \mathcal{N}$. \square

It follows from Proposition 2 that $[\mathbf{P}, \mathbf{C}]$ is a stable feedback configuration if and only if $\mathbf{A}_{\mathcal{M}, \mathcal{N}}$ is invertible. When $[\mathbf{P}, \mathbf{C}]$ is stable we define the operator

$$\mathbf{Q}_{\mathcal{M}, \mathcal{N}} := \mathbf{A}_{\mathcal{M}, \mathcal{N}}^{-1} \Pi_{\mathcal{N}^\perp}.$$

This is the *parallel projection onto \mathcal{M} along \mathcal{N}* . Note that $\mathbf{Q}_{\mathcal{M}, \mathcal{N}}$ can be expressed directly in terms of \mathbf{P} and \mathbf{C} as follows:

$$(23) \quad \mathbf{Q}_{\mathcal{M}, \mathcal{N}} = \begin{pmatrix} \mathbf{I}_\mathcal{U} \\ \mathbf{P} \end{pmatrix} ((\mathbf{I}_\mathcal{U} - \mathbf{C}\mathbf{P})^{-1}, -\mathbf{C}(\mathbf{I}_\mathcal{Y} - \mathbf{P}\mathbf{C})^{-1})$$

and in terms of the input-to-error operator $\mathbf{H}_{\mathbf{P}, \mathbf{C}}$:

$$(24) \quad \mathbf{Q}_{\mathcal{M}, \mathcal{N}} = \begin{pmatrix} \mathbf{I}_{\mathcal{U}} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{\mathcal{Y}} \end{pmatrix} \mathbf{H}_{\mathbf{P}, \mathbf{C}} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\mathcal{Y}} \end{pmatrix}.$$

When $[\mathbf{P}, \mathbf{C}]$ is stable we also define

$$(25) \quad \begin{aligned} b_{\mathcal{M}, \mathcal{N}} &:= \|\mathbf{Q}_{\mathcal{M}, \mathcal{N}}\|^{-1} \\ &= \tau(\Pi_{\mathcal{N}^\perp} |_{\mathcal{M}}) = \sqrt{1 - \delta(\mathcal{M}, \mathcal{N}^\perp)^2} \\ &= \inf\{\|\mathbf{A}_{\mathcal{M}, \mathcal{N}}x\| : x \in \mathcal{M} \text{ and } \|x\| = 1\} \\ &= \inf\{\text{dist}(x, \mathcal{N}) : x \in \mathcal{M} \text{ and } \|x\| = 1\} \\ &= \inf\{\sin \theta(x, y) : 0 \neq x \in \mathcal{M}, 0 \neq y \in \mathcal{N}\}, \end{aligned}$$

where

$$\theta(x, y) := \arccos \frac{|\langle x, y \rangle|}{\|x\| \|y\|}$$

denotes the *angle* between two nonzero vectors $x, y \in \mathcal{L}$. When $[\mathbf{P}, \mathbf{C}]$ is not stable we set $b_{\mathcal{M}, \mathcal{N}} := 0$. Equation (25) follows from (3). The quantity $b_{\mathcal{M}, \mathcal{N}}$ is the sine of the *minimal angle*

$$\begin{aligned} \theta_{\min}(\mathcal{M}, \mathcal{N}) &:= \inf\{\theta(x, y) : 0 \neq x \in \mathcal{M}, 0 \neq y \in \mathcal{N}\} \\ &= \arcsin b_{\mathcal{M}, \mathcal{N}} \\ &= \arccos \delta(\mathcal{M}, \mathcal{N}^\perp) \end{aligned}$$

(e.g., see [8]). Since

$$\begin{aligned} \delta(\mathcal{M}, \mathcal{N}^\perp) &= \delta(\mathcal{M}^\perp, \mathcal{N}) \\ &= \delta(\mathcal{N}, \mathcal{M}^\perp) \\ &= \delta(\mathcal{N}^\perp, \mathcal{M}), \end{aligned}$$

it follows that

$$(26) \quad \begin{aligned} b_{\mathcal{M}, \mathcal{N}} &= b_{\mathcal{M}^\perp, \mathcal{N}^\perp} \\ &= b_{\mathcal{N}, \mathcal{M}} \\ &= b_{\mathcal{N}^\perp, \mathcal{M}^\perp}. \end{aligned}$$

Equation (26) was shown in [3] (cf. [21, Lemma 1.1, p. 341]), [6]. It also follows from [10, Lemma 4] after noting that $b_{\mathcal{M}, \mathcal{N}}$ is the inverse of the norm of a parallel projection.

Conditions for stability of a feedback configuration can be expressed in a number of equivalent ways (cf. [3], [14]).

COROLLARY 3. *The following are equivalent:*

- (a) $[\mathbf{P}, \mathbf{C}]$ is stable;
- (b) $\mathbf{A}_{\mathcal{M}, \mathcal{N}}$ is invertible;
- (b) $\delta(\mathcal{M}, \mathcal{N}^\perp) < 1$;
- (c) $\theta_{\max}(\mathcal{M}, \mathcal{N}^\perp) < \pi/2$;
- (d) $\theta_{\min}(\mathcal{M}, \mathcal{N}) > 0$.

Proof. The proof follows from Propositions 3 and 6. \square

THEOREM 3. *The following are equivalent:*

- (a) $[\mathbf{P}, \mathbf{C}]$ is stable and $b < b_{\mathcal{M}, \mathcal{N}}$;
- (b) $[\mathbf{P}', \mathbf{C}]$ is stable and $\mathbf{Q}_{\mathcal{M}', \mathcal{N}}$ is uniformly bounded for all \mathbf{P}' so that, with $\mathcal{M}' := \mathcal{G}_{\mathbf{P}'}$, $\delta(\mathcal{M}, \mathcal{M}') \leq b$.

Before presenting the proof of the theorem we will establish the following lemma. For any $0 \neq h \in \mathcal{L}$ we denote $\hat{h} := h/\|h\|$.

LEMMA 3. *Let $\mathcal{K}, \mathcal{W} \in \mathcal{S}_{\mathcal{L}}$. Take any $0 \neq h \in \mathcal{K}$ and $0 \neq h_1 \notin \mathcal{K}$. Define $\mathcal{K}_- = \mathcal{K} \ominus \mathbb{C}h$ and $\mathcal{K}_1 = \mathcal{K}_- + \mathbb{C}h_1$. Then*

$$(a) \quad \delta(\mathcal{K}, \mathcal{K}_1) = \sqrt{1 - |\langle \hat{h}, \hat{h}_2 \rangle|^2}, \text{ where } h_2 = \Pi_{\mathcal{K}_-^\perp} h_1;$$

$$(b) \quad \text{if } \mathcal{K} \in \text{Graph}_{\mathcal{W}} \text{ then } \mathcal{K}_1 \in \overline{\text{Graph}_{\mathcal{W}}}.$$

Proof. (a) Since $\mathcal{K}_1 = \mathcal{K}_- \oplus \mathbb{C}h_2$ and $\mathcal{K}^\perp = \mathcal{K}_-^\perp \ominus \mathbb{C}h$ we obtain

$$\begin{aligned} \|\Pi_{\mathcal{K}^\perp} \Pi_{\mathcal{K}_1}\| &= \|(\Pi_{\mathcal{K}_-^\perp} - \Pi_{\mathbb{C}h})(\Pi_{\mathcal{K}_-} + \Pi_{\mathbb{C}h_2})\| \\ &= \|(\Pi_{\mathbb{C}h_2} - \Pi_{\mathbb{C}h}) \Pi_{\mathbb{C}h_2}\| \\ &= \sqrt{1 - |\langle \hat{h}, \hat{h}_2 \rangle|^2}. \end{aligned}$$

Similarly, $\|\Pi_{\mathcal{K}_1^\perp} \Pi_{\mathcal{K}}\| = \|(\Pi_{\mathbb{C}h_2} - \Pi_{\mathbb{C}h}) \Pi_{\mathbb{C}h}\| = \sqrt{1 - |\langle \hat{h}, \hat{h}_2 \rangle|^2}$. Therefore, $\delta(\mathcal{K}, \mathcal{K}_1) = \sqrt{1 - |\langle \hat{h}, \hat{h}_2 \rangle|^2}$.

(b) Define $\mathbf{X} := \Pi_{\mathcal{W}}|_{\mathcal{K}}$, $\mathbf{X}_1 := \Pi_{\mathcal{W}}|_{\mathcal{K}_1}$, $\hat{\mathbf{X}} := \mathbf{X} \Pi_{\mathcal{K}}|_{(\mathcal{K} + \mathcal{K}_1)}$, and $\hat{\mathbf{X}}_1 := \mathbf{X}_1 \Pi_{\mathcal{K}_1}|_{(\mathcal{K} + \mathcal{K}_1)}$. If $\mathcal{K}_1 \notin \overline{\text{Graph}_{\mathcal{W}}} = \mathcal{S}_{\text{npi}}$, then \mathbf{X}_1 is semi-Fredholm and then it is obvious that $\mathbf{X}, \hat{\mathbf{X}}, \hat{\mathbf{X}}_1$ are also semi-Fredholm. Since $\mathcal{K} \in \text{Graph}_{\mathcal{W}}$, $\ker \mathbf{X} = 0$ and $\text{ind } \mathbf{X} \leq 0$. Because $h_1 \notin \mathcal{K}$, it can be seen that $\dim \ker \hat{\mathbf{X}} = \dim \ker \mathbf{X} + 1$ and similarly that $\dim \ker \hat{\mathbf{X}}_1 = \dim \ker \mathbf{X}_1 + 1$. It follows that $\text{ind } \hat{\mathbf{X}} = \text{ind } \mathbf{X} + 1$ and $\text{ind } \hat{\mathbf{X}}_1 = \text{ind } \mathbf{X}_1 + 1$. However,

$$\hat{\mathbf{X}}_1 = \Pi_{\mathcal{W}} \Pi_{\mathcal{K}}|_{(\mathcal{K} + \mathcal{K}_1)} + \Pi_{\mathcal{W}}(\Pi_{\mathbb{C}h_2} - \Pi_{\mathbb{C}h})|_{(\mathcal{K} + \mathcal{K}_1)} = \hat{\mathbf{X}} + \text{finite-rank operator}.$$

Therefore, $\text{ind } \hat{\mathbf{X}}_1 = \text{ind } \hat{\mathbf{X}}$ and hence $\text{ind } \mathbf{X}_1 = \text{ind } \mathbf{X}$. Thus, $\text{ind } \mathbf{X}_1 \leq 0$, and consequently, $\mathcal{K}_1 \notin \mathcal{S}_{\text{pi}}$; that is, $\mathcal{K}_1 \in \mathcal{S}_{\text{npi}}$, a contradiction. This proves that $\mathcal{K}_1 \in \overline{\text{Graph}_{\mathcal{W}}}$. \square

Proof of Theorem 3. (a) \Rightarrow (b). Assume (b) fails. We will show that, if $[\mathbf{P}, \mathbf{C}]$ is stable, then $b \geq b_{\mathcal{M}, \mathcal{N}}$. Since (b) fails then, either $\|(\Pi_{\mathcal{N}^\perp}|_{\mathcal{M}'})^{-1}\|$ is not bounded above in $\overline{\text{Ball}}(\mathcal{M}, b) \cap \text{Graph}_{\mathcal{U}}$ or $\Pi_{\mathcal{N}^\perp}|_{\mathcal{M}'}$ is not invertible for some $\mathcal{M}' \in \overline{\text{Ball}}(\mathcal{M}, b) \cap \text{Graph}_{\mathcal{U}}$. This means that one of the following two possibilities holds:

- (i) $\tau(\Pi_{\mathcal{N}^\perp}|_{\mathcal{M}'})$ is not bounded below in $\overline{\text{Ball}}(\mathcal{M}, b) \cap \text{Graph}_{\mathcal{U}}$;
- (ii) there exists $\mathcal{M}' \in \overline{\text{Ball}}(\mathcal{M}, b) \cap \text{Graph}_{\mathcal{U}}$ and $0 \neq y \in \mathcal{N}^\perp$ such that $\Pi_{\mathcal{M}'} y = 0$.

In case (i), for all $\epsilon > 0$ there exists \mathcal{M}' and $x \in \mathcal{M}'$ of unit norm such that $\|\Pi_{\mathcal{N}^\perp} x\| = \text{dist}(x, \mathcal{N}) < \epsilon$. Note that $b \geq \delta(\mathcal{M}, \mathcal{M}') \geq \sup_{\xi \in \mathcal{M}', \|\xi\|=1} \text{dist}(\xi, \mathcal{M}) \geq \text{dist}(x, \mathcal{M}) = \|\Pi_{\mathcal{M}^\perp} x\|$. Also

$$\begin{aligned} b_{\mathcal{M}, \mathcal{N}} &= b_{\mathcal{N}, \mathcal{M}} \leq \left\| \Pi_{\mathcal{M}^\perp} \frac{\Pi_{\mathcal{N}} x}{\|\Pi_{\mathcal{N}} x\|} \right\| \\ (27) \quad &\leq \frac{\|\Pi_{\mathcal{M}^\perp} x\| + \|\Pi_{\mathcal{M}^\perp} \Pi_{\mathcal{N}^\perp} x\|}{\|\Pi_{\mathcal{N}} x\|} \leq \frac{b + \epsilon}{\sqrt{1 - \epsilon^2}}. \end{aligned}$$

Since (27) holds for all ϵ then $b_{\mathcal{M},\mathcal{N}} \leq b$. In case (ii), note that $y \in \mathcal{M}'^\perp$. Thus $b \geq \delta(\mathcal{M}, \mathcal{M}') = \delta(\mathcal{M}^\perp, \mathcal{M}'^\perp) \geq \text{dist}(y, \mathcal{M}^\perp) = \|\Pi_{\mathcal{M}} y\|$. Also $b_{\mathcal{M},\mathcal{N}} = b_{\mathcal{M}^\perp, \mathcal{N}^\perp} \leq \|\Pi_{\mathcal{M}} y\|$ since $y \in \mathcal{N}^\perp$. Therefore $b_{\mathcal{M},\mathcal{N}} \leq b$.

(b) \Rightarrow (a). Suppose that (b) holds for some $b \geq b_{\mathcal{M},\mathcal{N}}$. Then the same is true for some $b > b_{\mathcal{M},\mathcal{N}}$. To see this, first note that $b < 1$ necessarily, otherwise $[\mathbf{P}', \mathbf{C}]$ is stable for *any* system \mathbf{P}' , and there is an easy contradiction. By assumption, there exists a c such that $\|\mathbf{Q}_{\mathcal{M}',\mathcal{N}}\| \leq c$ for all \mathbf{P}' with $\delta(\mathcal{M}, \mathcal{M}') \leq b$. From the identity $\mathbf{A}_{\mathcal{M}'',\mathcal{N}} = \mathbf{A}_{\mathcal{M}',\mathcal{N}}(\mathbf{I}_{\mathcal{M}'} + \mathbf{Q}_{\mathcal{M}',\mathcal{N}}(\Pi_{\mathcal{M}''} - \Pi_{\mathcal{M}'})|_{\mathcal{M}'})(\Pi_{\mathcal{M}''}|_{\mathcal{M}'})^{-1}|_{\mathcal{M}''}$ (cf. [3]) we can see that $\mathbf{A}_{\mathcal{M}'',\mathcal{N}}$ is invertible for all \mathbf{P}'' with $\delta(\mathcal{M}', \mathcal{M}'') \leq 1/2c$ for some \mathbf{P}' with $\delta(\mathcal{M}, \mathcal{M}') \leq b$. Moreover

$$\begin{aligned} \|\mathbf{Q}_{\mathcal{M}'',\mathcal{N}}\| &= \|\Pi_{\mathcal{M}''}|_{\mathcal{M}'}(\mathbf{I}_{\mathcal{M}'} + \mathbf{Q}_{\mathcal{M}',\mathcal{N}}(\Pi_{\mathcal{M}''} - \Pi_{\mathcal{M}'})|_{\mathcal{M}'})^{-1}\mathbf{Q}_{\mathcal{M}',\mathcal{N}}\| \\ &\leq 2c. \end{aligned}$$

From Theorem 1 the union of open balls, of radius $1/2c$, about all \mathcal{M}' with $\delta(\mathcal{M}, \mathcal{M}') \leq b$ includes an open ball about \mathcal{M} of radius $b + \epsilon$ for some $\epsilon > 0$. It now follows that (b) holds for \mathbf{P}'' in a ball about \mathbf{P} and radius strictly greater than $b_{\mathcal{M},\mathcal{N}}$. So from now on we assume that $b_{\mathcal{M},\mathcal{N}} < b < 1$.

Next we prove that there exists a subspace $\mathcal{M}' \in \overline{\text{Graph}}_{\mathcal{U}}$ with $\delta(\mathcal{M}, \mathcal{M}') < b$ such that $\Pi_{\mathcal{N}^\perp}|_{\mathcal{M}'}$ is not invertible. Let $\mathbf{A} := \mathbf{A}_{\mathcal{M},\mathcal{N}}$, \mathbf{E}_λ be the spectral family of $\mathbf{A}^* \mathbf{A}$, and $h \in \mathbf{E}_{\tau(\mathbf{A})+\epsilon} \mathcal{M}$ of unit norm, for some arbitrary $\epsilon > 0$. Then $\|(\mathbf{A}^* \mathbf{A} - b_{\mathcal{M},\mathcal{N}})h\| \leq \epsilon$. Define $\mathcal{M}_- := \mathcal{M} \ominus \mathbb{C}h$, $p_0 := \Pi_{\mathcal{N}} h \in \mathcal{N}$, $q_0 := \Pi_{\mathcal{M}_-} p_0$, and $\mathcal{M}' = \mathcal{M}_- + \mathbb{C}p_0$. Since $\mathcal{M}' \cap \mathcal{N} \neq \{0\}$ then $\Pi_{\mathcal{M}'}|_{\mathcal{N}^\perp}$ is not invertible. From Lemma 3 we have

$$(28) \quad \delta(\mathcal{M}, \mathcal{M}') = \sqrt{1 - |\langle h, \hat{q}_0 \rangle|^2}$$

and $\mathcal{M}' \in \overline{\text{Graph}}_{\mathcal{U}}$. To evaluate (28) we first note that

$$(29) \quad \begin{aligned} \langle h, q_0 \rangle &= \langle h, p_0 \rangle = \|p_0\|^2 \\ &= 1 - \|\mathbf{A}h\|^2. \end{aligned}$$

We also have

$$(30) \quad \begin{aligned} \|q_0\|^2 &= \|p_0\|^2 - \|\Pi_{\mathcal{M}_-} \Pi_{\mathcal{N}} h\|^2 \\ &= \|p_0\|^2 - \|\Pi_{\mathcal{M}_-} \Pi_{\mathcal{N}^\perp} h\|^2 \\ &= \|p_0\|^2 - \langle \Pi_{\mathcal{N}^\perp} h, (\Pi_{\mathcal{M}} - \Pi_{\mathbb{C}h}) \Pi_{\mathcal{N}^\perp} h \rangle \\ &= \|p_0\|^2 - \langle \Pi_{\mathcal{M}} \Pi_{\mathcal{N}^\perp} h, \Pi_{\mathcal{M}} \Pi_{\mathcal{N}^\perp} h \rangle + |\langle \Pi_{\mathcal{N}^\perp} h, h \rangle|^2 \\ &= 1 - \|\mathbf{A}h\|^2 - \|\mathbf{A}^* \mathbf{A} h\|^2 + \|\mathbf{A}h\|^4 \\ &= 1 - \|\mathbf{A}h\|^2 + O(\epsilon). \end{aligned}$$

From (28)–(30), we obtain $\delta(\mathcal{M}, \mathcal{M}') = \|\mathbf{A}h\| + O(\epsilon)$. In particular, for sufficiently small ϵ we have $\delta(\mathcal{M}, \mathcal{M}') < b$.

In case $\mathcal{M}' \in \overline{\text{Graph}}_{\mathcal{U}}$ then the hypothesis is violated and the proof is complete. If not, consider a sequence $\mathcal{M}'_i \in \text{Graph}_{\mathcal{U}}$, $i = 1, 2, \dots$, converging to \mathcal{M}' . If there is a subsequence such that $\Pi_{\mathcal{M}'_i}|_{\mathcal{N}^\perp}$ is not invertible then, again, there is a contradiction. Otherwise, since $\lim_{i \rightarrow \infty} \|\Pi_{\mathcal{M}'_i}|_{\mathcal{N}^\perp} - \Pi_{\mathcal{M}'}|_{\mathcal{N}^\perp}\| = 0$, we can find a subsequence such that $\lim_{i \rightarrow \infty} \|(\Pi_{\mathcal{M}'_i}|_{\mathcal{N}^\perp})^{-1}\| = \infty$. This also violates the hypothesis. \square

Remark 1. A similar result was established in [6, Thm. 5] for linear time-invariant causal systems. However, the result in [6] differs from the one above in that the $<$

and \leq signs are interchanged, and the “uniform boundedness” is absent. We will show that the exact statement of [6, Thm. 5] is not valid in the case of time-varying systems so that the uniformity condition is in fact necessary. More precisely, in the next section we will present an example where $[\mathbf{P}_1, \mathbf{C}]$ is stable for all \mathbf{P}_1 such that $\delta(\mathcal{M}, \mathcal{M}_1) < 1$ while $b_{\mathcal{M}, \mathcal{N}} = 1/\sqrt{2} < 1$.

Remark 2. The basic geometric ideas behind the proof of Theorem 3 can be simply expressed. The essence of the sufficiency part of Theorem 3 ((a) \Rightarrow (b)) can be seen from the identity $\mathbf{A}_{\mathcal{M}', \mathcal{N}} = \mathbf{A}_{\mathcal{M}, \mathcal{N}}(\mathbf{I}_{\mathcal{M}} + \mathbf{Q}_{\mathcal{M}, \mathcal{N}}(\Pi_{\mathcal{M}'} - \Pi_{\mathcal{M}})|_{\mathcal{M}})(\Pi_{\mathcal{M}'}|_{\mathcal{M}})^{-1}|_{\mathcal{M}'}$, which implies that $\mathbf{A}_{\mathcal{M}', \mathcal{N}}$ is invertible for all \mathbf{P}' if $\delta(\mathcal{M}, \mathcal{M}') \leq \|\mathbf{Q}_{\mathcal{M}, \mathcal{N}}\|^{-1}$. The key idea in the necessity part is to find a subspace \mathcal{M}' such that $\delta(\mathcal{M}, \mathcal{M}') = \|\mathbf{Q}_{\mathcal{M}, \mathcal{N}}\|^{-1}$ and so that Proposition 6(a) is violated. The construction given in the proof is to remove a direction orthogonally from \mathcal{M} and to replace it with a direction from \mathcal{N} . The vectors are chosen in such a way that the angle between them is equal (or arbitrarily close to) $\theta_{\min}(\mathcal{M}, \mathcal{N})$. This gives $\mathcal{M}' \cap \mathcal{N} \neq \{0\}$ with $\delta(\mathcal{M}, \mathcal{M}')$ having the required value. (The reader is referred to [19] for another version of this idea.) The additional ingredients in the proof deal with the uniform boundedness condition and the need to impose graphability on the perturbed subspaces.

Remark 3. In the theorem we do not impose any time-invariance and causality constraint on the systems considered. Certainly the implication (a) \Rightarrow (b) of Theorem 3 is still valid when the class of systems is restricted by a causality requirement, but the reverse implication requires a construction different from the one given here.

6. Clarification of uniform boundedness condition. We now present an example to show that, in the time-varying case, the obstruction that limits the largest perturbation ball in the gap metric may be due solely to the lack of uniform boundedness of the closed loop operator, as expressed in Theorem 3.

Let $\mathcal{U} = \mathcal{Y} = \ell_2[0, \infty) =: \mathcal{V}$, $\mathcal{L} = \mathcal{U} \oplus \mathcal{Y}$, and identify \mathcal{U} and \mathcal{Y} with the corresponding subspaces of \mathcal{L} . Consider \mathbf{P} having the matrix representation

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & \dots \\ 0 & 0 & \dots \\ \vdots & \vdots & \end{pmatrix}$$

and let $\mathbf{C} = 0$. Then $\mathcal{M} = \{(\begin{smallmatrix} v \\ (v)_0 \end{smallmatrix}) : v \in \mathcal{V}\}$, where $(v)_0 = (v_0, 0, 0, \dots)$ for any $v = (v_0, v_1, v_2, \dots) \in \mathcal{V}$, and $\mathcal{N} = \{(\begin{smallmatrix} 0 \\ v \end{smallmatrix}) : v \in \mathcal{V}\}$. For any $\mathcal{M}_1 \in \text{Graph}_{\mathcal{U}}$ define \mathbf{P}_1 by $\mathcal{M}_1 = \mathcal{G}_{\mathbf{P}_1}$. Also define $\text{Ball}_{\mathcal{U}}(\mathcal{M}, b) := \text{Ball}(\mathcal{M}, b) \cap \text{Graph}_{\mathcal{U}}$.

PROPOSITION 7. *For the example given above*

$$\sup \{b : \{\mathcal{M}_1 \in \text{Ball}_{\mathcal{U}}(\mathcal{M}, b) \text{ implies that } [\mathbf{P}_1, \mathbf{C}] \text{ is stable}\} = 1.$$

This should be contrasted against the fact that for the particular \mathbf{P}, \mathbf{C} given above

$$\begin{aligned} b_{\mathcal{M}, \mathcal{N}} &= \left\| \begin{pmatrix} \mathbf{I}_{\mathcal{U}} \\ \mathbf{P} \end{pmatrix} ((\mathbf{I}_{\mathcal{U}} - \mathbf{C}\mathbf{P})^{-1}, -\mathbf{C}(\mathbf{I}_{\mathcal{Y}} - \mathbf{P}\mathbf{C})^{-1}) \right\|^{-1} \\ &= \left\| \begin{pmatrix} \mathbf{I}_{\mathcal{U}} \\ \mathbf{P} \end{pmatrix} \right\|^{-1} = (1 + \|\mathbf{P}\|^2)^{-1/2} = \frac{1}{\sqrt{2}}. \end{aligned}$$

Proof of Proposition 7. Since $\mathcal{N} = \mathcal{U}^{\perp}$, for any $\mathcal{M}_1 \in \text{Ball}_{\mathcal{U}}(\mathcal{M}, b)$, $\mathcal{M}_1 \cap \mathcal{N} = \{0\}$. Therefore, $[\mathbf{P}_1, \mathbf{C}]$ is stable $\Leftrightarrow \mathcal{M}_1 + \mathcal{N} = \mathcal{L} \Leftrightarrow \Pi_{\mathcal{U}}\mathcal{M}_1 = \mathcal{U}$. Next note

that $\mathcal{M}^\perp = \left\{ \begin{pmatrix} -(v)_0 \\ v \end{pmatrix} : v \in \mathcal{V} \right\}$. Proposition 3 also implies that any \mathcal{M}_1 such that $\delta(\mathcal{M}, \mathcal{M}_1) < 1$ can be written as

$$\mathcal{M}_1 = \left\{ \begin{pmatrix} v \\ (v)_0 \end{pmatrix} + \begin{pmatrix} -(\mathbf{X}v)_0 \\ \mathbf{X}v \end{pmatrix} : v \in \mathcal{V} \right\},$$

where $\mathbf{X} : \mathcal{V} \rightarrow \mathcal{V}$ is a bounded operator. However,

$$\begin{aligned} \mathcal{M}_1 \cap \mathcal{N} = \{0\} &\Leftrightarrow v - (\mathbf{X}v)_0 = 0 \text{ implies } (v)_0 + \mathbf{X}v = 0 \\ &\Leftrightarrow v = (v)_0 = (\mathbf{X}v)_0 \text{ implies } (v)_0 + \mathbf{X}v = 0 \\ &\Leftrightarrow (\mathbf{X})_{0,0} \neq 1, \end{aligned}$$

where $(\mathbf{X})_{0,0}$ denotes the $(0,0)$ -entry in a matrix representation of \mathbf{X} with respect to the standard basis of $\mathcal{V} = \ell_2[0, \infty)$. Take any $\mathcal{M}_1 \in \text{Ball}_{\mathcal{U}}(\mathcal{M}, 1)$. Then

$$\Pi_{\mathcal{U}}\mathcal{M}_1 = \{v - (\mathbf{X}v)_0 : v \in \mathcal{V}\} = \mathcal{U}$$

since $(\mathbf{X})_{0,0} \neq 1$. Hence, $\mathcal{M}_1 \in \text{Ball}_{\mathcal{U}}(\mathcal{M}, 1)$ implies that $[\mathbf{P}_1, \mathbf{C}]$ is stable. So $b = 1$ is the supremal b . \square

7. Combined plant and controller uncertainty. When both plant and controller are subject simultaneously to gap-ball uncertainty, there is a maximal amount for the combined uncertainty that can be tolerated. The following theorem is a generalization of an elegant result of Qiu and Davison [16] to the time-varying case.

THEOREM 4. *Let $\mathbf{P} \in \mathcal{P}_{\mathcal{U}, \mathcal{Y}}$, $\mathbf{C} \in \mathcal{P}_{\mathcal{Y}, \mathcal{U}}$ and let b_1, b_2 be fixed nonnegative numbers such that $b_1^2 + b_2^2 < 1$. Then the following are equivalent:*

- (a) $[\mathbf{P}, \mathbf{C}]$ is stable and $b_1\sqrt{1-b_2^2} + b_2\sqrt{1-b_1^2} < b_{\mathcal{M}, \mathcal{N}}$;
- (b) $[\mathbf{P}', \mathbf{C}']$ is stable and $\mathbf{Q}_{\mathcal{M}', \mathcal{N}'}$ is uniformly bounded for all \mathbf{P}', \mathbf{C}' with $\mathcal{M}' := \mathcal{G}_{\mathbf{P}'}, \mathcal{N}' := \mathcal{G}_{\mathbf{C}'}, \delta(\mathcal{M}, \mathcal{M}') \leq b_1$ and $\delta(\mathcal{N}, \mathcal{N}') \leq b_2$.

Proof. (a) \Rightarrow (b) Suppose (b) fails and that $[\mathbf{P}, \mathbf{C}]$ is stable. We will show that $b_1\sqrt{1-b_2^2} + b_2\sqrt{1-b_1^2} \geq b_{\mathcal{M}, \mathcal{N}}$. As in the proof of Theorem 3 there are two possibilities:

(i) $\tau(\Pi_{\mathcal{N}'^\perp} |_{\mathcal{M}'})$ is not bounded below for $\mathcal{M}' \in \overline{\text{Ball}}(\mathcal{M}, b_1) \cap \text{Graph}_{\mathcal{U}}$ and $\mathcal{N}' \in \overline{\text{Ball}}(\mathcal{N}, b_2) \cap \text{Graph}_{\mathcal{Y}}$,

(ii) there exists $\mathcal{M}' \in \overline{\text{Ball}}(\mathcal{M}, b_1) \cap \text{Graph}_{\mathcal{U}}$ and $\mathcal{N}' \in \overline{\text{Ball}}(\mathcal{N}, b_2) \cap \text{Graph}_{\mathcal{Y}}$ and $0 \neq z \in \mathcal{N}'^\perp \cap \mathcal{M}'^\perp$.

In case (i), for all $\epsilon > 0$ there exists $\mathcal{M}', \mathcal{N}'$, and $x \in \mathcal{M}'$ of unit norm such that $\|\Pi_{\mathcal{N}'^\perp} x\| < \epsilon$. Setting $y := \Pi_{\mathcal{N}'} x \in \mathcal{N}'$ we have

$$\theta(x, y) := \arccos \frac{|\langle x, y \rangle|}{\|x\| \|y\|} = \arccos \left(\frac{1 - \|\Pi_{\mathcal{N}'^\perp} x\|^2}{\|y\|} \right) < \arcsin \epsilon.$$

Since $\delta(\mathcal{M}, \mathcal{M}') \leq b_1$ it follows that $\|\Pi_{\mathcal{M}^\perp} x\| \leq b_1$. Thus, if $x_0 := \Pi_{\mathcal{M}} x \in \mathcal{M}$, then

$$\theta(x_0, x) \leq \arcsin b_1.$$

Similarly, since $\delta(\mathcal{N}, \mathcal{N}') \leq b_2$,

$$\theta(y_0, y) \leq \arcsin b_2,$$

where $y_0 := \Pi_{\mathcal{N}} y \in \mathcal{N}$. It follows from (15) that

$$\begin{aligned} \arcsin b_1 + \arcsin b_2 + \arcsin \epsilon &> \theta(x_0, x) + \theta(y, y_0) + \theta(x, y) \\ &\geq \theta(x_0, y_0) \\ (31) \quad &\geq \theta_{\min}(\mathcal{M}, \mathcal{N}) = \arcsin b_{\mathcal{M}, \mathcal{N}}. \end{aligned}$$

Since (31) holds for all ϵ we have

$$\arcsin b_1 + \arcsin b_2 \geq \arcsin b_{\mathcal{M}, \mathcal{N}}.$$

Therefore, $b_1 \sqrt{1 - b_2^2} + b_2 \sqrt{1 - b_1^2} \geq b_{\mathcal{M}, \mathcal{N}}$ and so (a) fails. In case (ii) we proceed similarly. Set $z_1 := \Pi_{\mathcal{M}^\perp} z$ and $z_2 := \Pi_{\mathcal{N}^\perp} z$. Since $b_1 \geq \delta(\mathcal{M}, \mathcal{M}') = \delta(\mathcal{M}^\perp, \mathcal{M}'^\perp)$ we have $\theta(z, z_1) \leq \arcsin b_1$. Also, $b_2 \geq \delta(\mathcal{N}, \mathcal{N}') = \delta(\mathcal{N}^\perp, \mathcal{N}'^\perp)$ implies that $\theta(z, z_2) \leq \arcsin b_2$. Thus

$$\begin{aligned} \arcsin b_1 + \arcsin b_2 &\geq \theta(z_1, z_2) \\ &\geq \theta_{\min}(\mathcal{M}^\perp, \mathcal{N}^\perp) \\ &= \arcsin b_{\mathcal{M}^\perp, \mathcal{N}^\perp} = \arcsin b_{\mathcal{M}, \mathcal{N}} \end{aligned}$$

and (a) fails once again.

(b) \Rightarrow (a) Suppose (b) holds for some b_1 and b_2 satisfying

$$(32) \quad b_1 \sqrt{1 - b_2^2} + b_2 \sqrt{1 - b_1^2} \geq b_{\mathcal{M}, \mathcal{N}}.$$

Similar reasoning to the proof of Theorem 3 shows that we may take strict inequality in (32). In particular, if $\|\mathbf{Q}_{\mathcal{M}', \mathcal{N}'}\| \leq c$ in (b), then $\|\mathbf{Q}_{\mathcal{M}'', \mathcal{N}'}\| \leq 2c$ for all

$$\mathcal{M}'' \in \bigcup_{\mathcal{M}' \in \overline{\text{Ball}}_{\mathcal{U}}(\mathcal{M}, b_1)} \text{Ball}_{\mathcal{U}}\left(\mathcal{M}', \frac{1}{2c}\right)$$

and all $\mathcal{N}' \in \text{Ball}_{\mathcal{Y}}(\mathcal{N}, b_2)$. Theorem 1 then shows that statement (b) holds with b_1 replaced by some $b_1 + \epsilon$ for $\epsilon > 0$. Since the left-hand side of (32) is monotonically increasing in b_1 , it follows that (b) holds for some b_1 and b_2 satisfying (32) with strict inequality. Henceforth we will assume that this is the case. We also note from Theorem 3 that $b_1, b_2 < b_{\mathcal{M}, \mathcal{N}}$.

We now show that there are subspaces $\mathcal{M}' \in \overline{\text{Graph}}_{\mathcal{U}}$ and $\mathcal{N}' \in \overline{\text{Graph}}_{\mathcal{Y}}$ with $\delta(\mathcal{M}, \mathcal{M}') < b_1$ and $\delta(\mathcal{N}, \mathcal{N}') < b_2$ such that $\Pi_{\mathcal{N}'^\perp}|_{\mathcal{M}'}$ is not invertible.

As in the proof of Theorem 3, let $\mathbf{A} := \mathbf{A}_{\mathcal{M}, \mathcal{N}}$, \mathbf{E}_λ be the spectral family of $\mathbf{A}^* \mathbf{A}$, and $h \in \mathbf{E}_{\tau(\mathbf{A}) + \epsilon} \mathcal{M}$ of unit norm, for some arbitrary $\epsilon > 0$. Then $\|\mathbf{A}h\| < \tau(\mathbf{A}) + \epsilon = b_{\mathcal{M}, \mathcal{N}} + \epsilon$. Define $p_\lambda = \lambda h + (1 - \lambda)\Pi_{\mathcal{N}}h$ and write $\mathcal{M}_- := \mathcal{M} \ominus \mathbb{C}h$, $\mathcal{N}_- := \mathcal{N} \ominus \mathbb{C}\Pi_{\mathcal{N}}h$, $\mathcal{M}_\lambda := \mathcal{M}_- + \mathbb{C}p_\lambda$, and $\mathcal{N}_\lambda := \mathcal{N}_- + \mathbb{C}p_\lambda$. We also write $q_\lambda := \Pi_{\mathcal{M}_-^\perp} p_\lambda$ and $r_\lambda := \Pi_{\mathcal{N}_-^\perp} p_\lambda$. We first show that

$$(33) \quad \delta(\mathcal{M}, \mathcal{M}_\lambda) = \frac{(1 - \lambda)\|\mathbf{A}h\|\sqrt{1 - \|\mathbf{A}h\|^2}}{\sqrt{1 - (1 - \lambda^2)\|\mathbf{A}h\|^2}} + O(\epsilon) =: c_\lambda + O(\epsilon).$$

From Lemma 3 we know that

$$(34) \quad \delta(\mathcal{M}, \mathcal{M}_\lambda) = \sqrt{1 - |\langle h, \hat{q}_\lambda \rangle|^2}.$$

To evaluate (34) we must compute $\langle h, q_\lambda \rangle$ and $\|q_\lambda\|$. First,

$$\begin{aligned} \langle h, q_\lambda \rangle &= \langle h, p_\lambda \rangle \\ &= \lambda + (1 - \lambda)\|p_0\|^2 \\ &= \lambda + (1 - \lambda)(1 - \|\mathbf{A}h\|^2) \\ (35) \quad &= 1 - (1 - \lambda)\|\mathbf{A}h\|^2. \end{aligned}$$

Next, using (30), we have

$$\begin{aligned}
 \langle q_\lambda, q_\lambda \rangle &= \lambda^2 + 2\lambda(1-\lambda)\|p_0\|^2 + (1-\lambda)^2\|q_0\|^2 \\
 &\simeq \lambda^2 + (2\lambda(1-\lambda) + (1-\lambda)^2)(1 - \|\mathbf{A}h\|^2) \\
 (36) \qquad &= 1 - (1-\lambda^2)\|\mathbf{A}h\|^2,
 \end{aligned}$$

where \simeq denotes equality to $O(\epsilon)$. Equations (35) and (36) together show that

$$|\langle h, \hat{q}_\lambda \rangle|^2 = \frac{(1 - (1-\lambda)\|\mathbf{A}h\|^2)^2}{1 - (1-\lambda^2)\|\mathbf{A}h\|^2} + O(\epsilon)$$

from which (33) follows by simple manipulation. Next we show that

$$(37) \qquad \delta(\mathcal{N}, \mathcal{N}_\lambda) = \frac{\lambda\|\mathbf{A}h\|}{\sqrt{1 - (1-\lambda^2)\|\mathbf{A}h\|^2}} =: d_\lambda.$$

From Lemma 3 we know that

$$(38) \qquad \delta(\mathcal{N}, \mathcal{N}_\lambda) = \sqrt{1 - |\langle \hat{p}_0, \hat{r}_\lambda \rangle|^2}.$$

To evaluate (34) we need the following computations:

$$\begin{aligned}
 \langle p_0, r_\lambda \rangle &= \langle \Pi_{\mathcal{N}} h, (\Pi_{\mathcal{N}^\perp} + \Pi_{\mathcal{T}p_0}) p_\lambda \rangle \\
 &= \langle \Pi_{\mathcal{N}} h, h \rangle \\
 (39) \qquad &= 1 - \|\mathbf{A}h\|^2
 \end{aligned}$$

$$(40) \qquad = \|p_0\|^2$$

and

$$\begin{aligned}
 \langle r_\lambda, r_\lambda \rangle &= \langle (\Pi_{\mathcal{N}^\perp} + \Pi_{\mathcal{T}p_0}) p_\lambda, (\Pi_{\mathcal{N}^\perp} + \Pi_{\mathcal{T}p_0}) p_\lambda \rangle \\
 &= \|\Pi_{\mathcal{N}^\perp} p_\lambda\|^2 + \|\Pi_{\mathcal{T}p_0} p_\lambda\|^2 \\
 &= \lambda^2 \|\Pi_{\mathcal{N}^\perp} h\|^2 + \|\Pi_{\mathcal{N}} h\|^2 \\
 (41) \qquad &= 1 - (1-\lambda^2)\|\mathbf{A}h\|^2.
 \end{aligned}$$

Equations (39)–(41) together show that

$$|\langle \hat{p}_0, \hat{r}_\lambda \rangle|^2 = \frac{1 - \|\mathbf{A}h\|^2}{1 - (1-\lambda^2)\|\mathbf{A}h\|^2}$$

from which (37) follows by simple manipulation. Next we observe from (33) and (37) that $c_\lambda \sqrt{1 - d_\lambda^2} + d_\lambda \sqrt{1 - c_\lambda^2} = \|\mathbf{A}h\|$. Since c_λ is monotonically decreasing in λ on the interval $[0, 1]$ we can choose λ such that $c_\lambda = b_1 - \epsilon$. Then for sufficiently small ϵ , we must have $d_\lambda < b_2$; otherwise, we have a contradiction to (32) with strict inequality. For the above choice of λ and sufficiently small ϵ we set $\mathcal{M}' = \mathcal{M}_\lambda$ and $\mathcal{N}' = \mathcal{N}_\lambda$ which gives $\delta(\mathcal{M}, \mathcal{M}') < b_1$ and $\delta(\mathcal{N}, \mathcal{N}') < b_2$. Lemma 3 shows that $\mathcal{M}' \in \overline{\text{Graph}}_{\mathcal{U}}$ and $\mathcal{N}' \in \overline{\text{Graph}}_{\mathcal{Y}}$. Also $\Pi_{\mathcal{N}'^\perp}|_{\mathcal{M}'}$ is not invertible since $\mathcal{M}' \cap \mathcal{N}' \neq \{0\}$.

Now consider a sequence $\mathcal{M}'_i \in \text{Graph}_{\mathcal{U}}$, $i = 1, 2, \dots$, converging to \mathcal{M}' and a sequence $\mathcal{N}'_i \in \text{Graph}_{\mathcal{Y}}$, $i = 1, 2, \dots$, converging to \mathcal{N}' . If there is a subsequence such that $\Pi_{\mathcal{N}'_i^\perp}|_{\mathcal{M}'_i}$ is not invertible, then there is a contradiction. Otherwise, we can find a subsequence so that $\Pi_{\mathcal{N}'_i^\perp}|_{\mathcal{M}'_i}$ is invertible. First observe that $\lim_{i \rightarrow \infty} \|\Pi_{\mathcal{N}'^\perp} \Pi_{\mathcal{M}'} -$

$\Pi_{\mathcal{N}'^\perp} \Pi_{\mathcal{M}'_i} = 0$. Since $\Pi_{\mathcal{N}'^\perp}|_{\mathcal{M}'}$ is not invertible, for any ϵ we can find an $x \in \mathcal{M}'$ of unit norm such that $\|\Pi_{\mathcal{N}'^\perp} x\| < \epsilon$. Thus $\|\Pi_{\mathcal{N}'^\perp} y_i\| < \epsilon$ for sufficiently large i , where $y_i := \Pi_{\mathcal{M}'_i} x$. Since $\mathcal{M}'_i \rightarrow \mathcal{M}'$ we also have $\|\Pi_{\mathcal{N}'^\perp} \hat{y}_i\| < \epsilon$ for sufficiently large i . This means that $\lim_{i \rightarrow \infty} \|(\Pi_{\mathcal{N}'^\perp}|_{\mathcal{M}'_i})^{-1}\| = \infty$. This violates the hypothesis. \square

Remark. The necessity part of the proof of Theorem 4 requires a simultaneous perturbation of \mathcal{M} and \mathcal{N} . The construction removes orthogonally one-dimensional subspaces from each of \mathcal{M} and \mathcal{N} that are at an angle $\theta_{\min}(\mathcal{M}, \mathcal{N})$ to each other, and replaces them by a convex combination of these directions. The subspaces are each perturbed through the required minimal angles and together violate the direct sum property of Proposition 6(a).

Acknowledgments. The authors would like to thank the reviewers for several helpful comments.

REFERENCES

- [1] A. K. EL-SAKKARY, *The gap metric: Robustness of stabilization of feedback systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 240–247.
- [2] A. FEINTUCH, *The gap metric for time-varying systems*, Systems Control Lett., 16 (1991), pp. 277–279.
- [3] C. FOIAS, T. T. GEORGIOU, AND M. C. SMITH, *Geometric techniques for robust stabilization of linear time-varying systems*, preprint, February, 1990; Proceedings of the 1990 IEEE Conference on Decision and Control, Hawaii, December, 1990.
- [4] ———, *Robust stabilization in the gap metric: a geometric approach*, Proceedings of the 1991 International Symposium on the Mathematical Theory of Networks and Systems, Kobe, Japan, June, 1991.
- [5] T. T. GEORGIOU, *On the computation of the gap metric*, Systems and Control Lett., 11 (1988), pp. 253–257.
- [6] T. T. GEORGIOU AND M. C. SMITH, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.
- [7] K. GLOVER AND D. MCFARLANE, *Robust stabilization of normalized coprime factor plant descriptions with H_∞ bounded uncertainty*, IEEE Trans. Automat. Control, 34 (1989), pp. 821–830.
- [8] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.
- [9] P. R. HALMOS, *A Hilbert Space Problem Book*, 2nd Edition, Springer-Verlag, New York, 1982.
- [10] T. KATO, *Estimation of iterated matrices, with application to the von Neumann condition*, Numer. Math., 2 (1960), pp. 22–29.
- [11] ———, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [12] M. A. KRASNOSEL'SKII, G. M. VAINIKKO, P. P. ZABREIKO, YA. B. RUTITSKII, AND V. YA. STETSENKO, *Approximate Solution of Operator Equations*, Wolters-Noordhoff, Groningen, 1972.
- [13] M. G. KREĬN AND M. A. KRASNOSEL'SKII, *Fundamental theorems concerning the extension of Hermitian operators and some of their applications to the theory of orthogonal polynomials and the moment problem*, Uspekhi Mat. Nauk, 2 (1947), pp. 60–106. (In Russian.)
- [14] R. OBER AND J. SEFTON, *Stability of Linear Systems and Graphs of Linear Systems*, Technical Report # 202, The University of Texas at Dallas, August, 1990.
- [15] ———, *Stability of linear systems and graphs of linear systems*, Systems Control Lett., 17 (1991), pp. 265–280.
- [16] L. QIU AND E. J. DAVISON, *Feedback stability under simultaneous gap metric uncertainties in plant and controller*, Systems Control Lett., 18 (1992), pp. 9–22.
- [17] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Dover, 1990.
- [18] M. SAFONOV, *Stability and Robustness of Multivariable Feedback Systems*, MIT Press, Cambridge, MA, 1980.
- [19] J. M. SCHUMACHER, *A pointwise criterion for controller robustness*, Systems Control Lett., 18 (1992), pp. 1–8.
- [20] B. SZ.-NAGY, *Perturbations des transformations autoadjointes dans l'espace de Hilbert*, Comment. Math. Helv., 19 (1947), pp. 347–366.

- [21] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.
- [22] M. S. VERMA, *Coprime fractional representations and stability of non-linear feedback systems*, Internat. J. Control, 48 (1988), pp. 897–918.
- [23] M. VIDYASAGAR, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control., 29 (1984), pp. 403–418.
- [24] M. VIDYASAGAR AND H. KIMURA, *Robust controllers for uncertain linear multivariable systems*, Automatica, 22 (1986), pp. 85–94.
- [25] G. ZAMES AND A. K. EL-SAKKARY, *Unstable systems and feedback: The gap metric*, Proc. of the Allerton Conference, October, 1980, pp. 380–385.
- [26] S. Q. ZHU, M. L. J. HAUTUS, AND C. PRAAGMAN, *Sufficient conditions for robust BIBO stabilization: Given by the gap metric*, Systems Control Lett., 11 (1988), pp. 53–59.
- [27] S. Q. ZHU, *Graph topology and gap topology for unstable systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 848–855.

STABILIZABILITY AND EXISTENCE OF SYSTEM REPRESENTATIONS FOR DISCRETE-TIME TIME-VARYING SYSTEMS*

WILBUR N. DALE[†] AND MALCOLM C. SMITH[‡]

Abstract. In this paper, right and left representations as an alternate, but equivalent, framework to coprime factorizations of operators are developed. The main theorem of the paper establishes that a linear, time-varying, discrete-time plant is stabilizable if and only if its graph can be represented as the range (respectively, kernel) of a causal, bounded operator which is left (respectively, right) invertible. The proof relies on certain factorization theorems of Arveson for nest algebras. The paper extends the Youla parametrization of all stabilizing compensators to this framework. Also, it is proven that a time-invariant plant that is not stabilizable by a time-invariant compensator is not stabilizable with a time-varying compensator. An example of a time-varying plant of Feintuch is considered and shown to be not stabilizable. Finally, the continuous-time case is examined and the problems encountered in extending the proof are discussed. However, it is shown that a stabilizable plant must have a closed graph and this is used to prove that an example of a time-invariant, continuous-time system of Shefi is not stabilizable.

Key words. time-varying systems, stabilizability, coprime fractions, graph, nest algebras

AMS subject classifications. 93C05, 93C25, 93C50, 93C55, 93D25

1. Notation and definitions. In this section, we will introduce notation and definitions used throughout the paper.

Let \mathbb{Z} be the nonnegative integers and \mathbb{R} be the set of real numbers. Let $\mathbb{R}^{n \times m}$ be the set of real matrices of n rows by m columns with $\mathbb{R}^{n \times 1} = \mathbb{R}^n$ being the n th-dimensional real Hilbert space with the inner product $\langle x, y \rangle = x^T y$, where x^T denotes the transpose of x . The norm of an element of \mathbb{R}^n is $\|x\| = \langle x, x \rangle^{1/2}$.

The set of all square summable sequences of \mathbb{R}^n is $\ell_2^n[0, \infty)$. That is, $\{x_k\} \in \ell_2^n[0, \infty)$ if

$$\sum_{k=0}^{\infty} \|x_k\|^2 < \infty.$$

$\ell_2^n[0, \infty)$ is also a Hilbert space with the inner product

$$\langle \{x_k\}, \{y_k\} \rangle = \sum_{k=0}^{\infty} \langle x_k, y_k \rangle.$$

The norm of an element in ℓ_2^n is $\|\{x_k\}\| = \langle \{x_k\}, \{x_k\} \rangle^{1/2}$.

If \mathcal{S} is a subspace of ℓ_2^n , then $\bar{\mathcal{S}}$ denotes the closure and \mathcal{S}^\perp the orthogonal complement. We denote the domain of an operator $G : \ell_2^m \rightarrow \ell_2^n$ by $\mathcal{D}\{G\}$, its range by $\mathcal{R}\{G\}$, and its kernel by $\mathcal{K}\{G\}$. The operator is defined only on its domain and the domain need not be closed. Let G^* denote the adjoint of the operator G . The graph $\mathcal{G}\{G\}$ of an operator $G : \ell_2^m \rightarrow \ell_2^n$ is defined to be

$$\mathcal{G}\{G\} := \begin{bmatrix} I \\ G \end{bmatrix} \mathcal{D}\{G\}.$$

* Received by the editors January 28, 1991; accepted for publication (in revised form) April 24, 1992.

[†] Edison Welding Institute, Columbus, Ohio 43212.

[‡] Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom.

The *inverse graph* $\mathcal{G}^{-1}\{G\}$ of G is

$$\mathcal{G}^{-1}\{G\} := \begin{bmatrix} G \\ I \end{bmatrix} \mathcal{D}\{G\}.$$

An operator G is *bounded* with *norm* $\|G\|$ if $\mathcal{D}\{G\} = \ell_2^m$ and

$$\|G\| = \sup_{x \in \mathcal{D}\{G\}} \frac{\|Gx\|}{\|x\|} < \infty.$$

Let S_m denote the shift operator on ℓ_2^m

$$S_m \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ x_0 \\ x_1 \\ \vdots \end{bmatrix}$$

for $\{x_k\} \in \ell_2^m[0, \infty)$. An operator G is said to be *shift-invariant* or *time-invariant* if $S_{m+n}\mathcal{G}\{G\} \subset \mathcal{G}\{G\}$.

Let $P_m(k)$ denote the truncation operator on ℓ_2^m

$$P_m(k) \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{k-1} \\ x_k \\ x_{k+1} \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ x_k \\ x_{k+1} \\ \vdots \end{bmatrix}$$

for $\{x_k\} \in \ell_2^m[0, \infty)$. An operator G is said to be *causal* if

$$\mathcal{G}\{G\} \cap \begin{pmatrix} P_m(k)\ell_2^m \\ \ell_2^n \end{pmatrix} \subset P_{m+n}(k)\ell_2^{m+n}$$

for all $k \in \mathbb{Z}$. If G is bounded, this condition reduces to $[I - P_n(k)]GP_m(k) = 0$ for all $k \in \mathbb{Z}$.

Consider the feedback system in Fig. 1, which we denote by $\{G, F\}$. The operators $G: \ell_2^m \rightarrow \ell_2^n$ and $F: \ell_2^n \rightarrow \ell_2^m$ represent the plant and the compensator, respectively. The closed-loop system equations are

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} I & F \\ G & I \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}.$$

We will assume throughout that G and F are linear and causal, although possibly time-varying and unbounded.

DEFINITION 1.1. The closed-loop system $\{G, F\}$ is *stable* if

$$\begin{bmatrix} I & F \\ G & I \end{bmatrix} : \mathcal{D}\{G\} \times \mathcal{D}\{F\} \longrightarrow \ell_2^m \times \ell_2^n$$

has a bounded inverse defined on $\ell_2^m \times \ell_2^n$. A plant G is *stabilizable* if there exists a compensator F such that $\{G, F\}$ is stable.

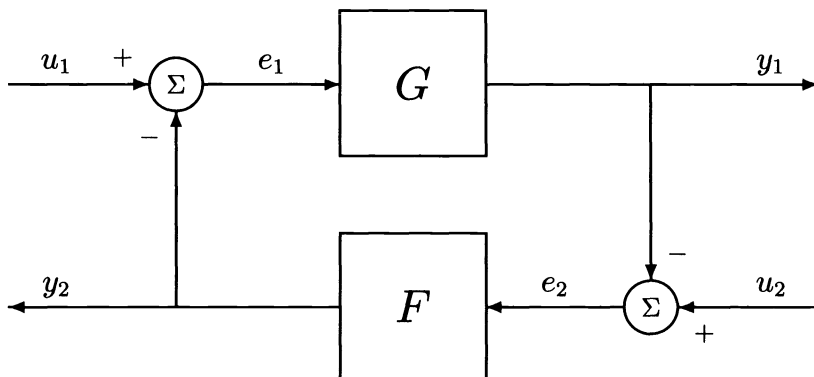


FIG. 1. Standard feedback configuration.

The following result was proved in [6] for the case of linear systems defined over $\ell_2[0, \infty)$.

THEOREM 1.2. *Suppose the closed-loop system $\{G, F\}$ is stable. Then the operator*

$$\begin{bmatrix} I & F \\ G & I \end{bmatrix}^{-1} : \ell_2^m \times \ell_2^n \longrightarrow \mathcal{D}\{G\} \times \mathcal{D}\{F\}$$

is causal.

Proof. Since the mapping

$$K := \begin{bmatrix} I & F \\ G & I \end{bmatrix}$$

is causal, it induces a well-defined linear map from a subspace of $\ell_2^{m+n}[0, k]$ onto $\ell_2^{m+n}[0, k]$ for all k . This mapping must have a matrix representation and the representation must be square and nonsingular, otherwise it cannot be onto. Now suppose that

$$H := \begin{bmatrix} I & F \\ G & I \end{bmatrix}^{-1}$$

is not causal. Then there exists $x, y \in \ell_2^{m+n}$ such that $y = Hx$, with $(I - P_{m+n}(k+1))x = 0$ and $(I - P_{m+n}(k+1))y \neq 0$ for some k . However, $x = Ky$, so the restriction of K to $[0, k]$ must have a kernel. This contradicts the fact that the matrix representation of this restriction is nonsingular. \square

2. Problem description. Coprime factorizations of linear, time-invariant systems have been extensively studied in recent years. The following theorem summarizes several results that have been obtained for the case of linear systems on $\ell_2[0, \infty)$ or $L_2[0, \infty)$ (the continuous-time Lebesgue 2-space). These results are expressed in terms of factorizations of the matrix transfer function over the space H_∞ (the standard Hardy space of the disk or right half plane, respectively). See [3], [18], [17], [10], [15] and the references therein for further details.

THEOREM 2.1. *A linear, time-invariant plant with transfer-function G (either continuous-time or discrete-time) is stabilizable if and only if there exists functions $M, N, X, Y, \widetilde{M}, \widetilde{N}, \widetilde{X}$, and $\widetilde{Y} \in H_\infty$ with*

$$(1) \quad G = NM^{-1} = \widetilde{M}^{-1}\widetilde{N}$$

such that the following double Bezout identity holds

$$(2) \quad \begin{bmatrix} Y & X \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \begin{bmatrix} M & -\widetilde{X} \\ N & \widetilde{Y} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Also, a linear, time-invariant compensator F stabilizes G if and only if

$$(3) \quad \begin{aligned} F &= (-\widetilde{X} - MQ) (\widetilde{Y} - NQ)^{-1} \\ &= (Y - Q\widetilde{N})^{-1} (-X - Q\widetilde{M}) \end{aligned}$$

for some $Q \in H_\infty$.

Several articles have appeared in the literature exploring similar factorizations for time-varying systems or nonlinear systems. Feintuch [4] found a necessary condition for the existence of coprime factorizations for discrete-time systems. Also, an example was presented of a plant that has no coprime factorization. Hammer [9] developed a framework for coprime factorizations for discrete-time, nonlinear systems on ℓ_∞ (see also [8]). Verma [16] examined fractional representations for nonlinear, time-varying systems in a general setting that included continuous-time systems. In both [9], [16], the existence of strong coprime factorizations for both the plant and the compensator was assumed. Thus far, the question of whether all stabilizable plants have strong coprime factorizations has remained unresolved. However, for the set of plants that can be realized by a finite set of state equations, such results do exist. Poolla and Khargonekar [12] examined linear, discrete-time, time-varying, finite-dimensional plants and proved the existence of coprime factorizations for stabilizable plants. Rotea and Khargonekar in [13] proved the existence of coprime factorizations for continuous-time, finite-dimensional, stabilizable plants. However, there are many infinite-dimensional (distributed parameter) plants that do not satisfy this criteria.

In this paper, we develop an alternate, but equivalent framework to the coprime factorization in previous work. In an operator-theoretic approach to coprime factorization, we work with products of operators in the form NM^{-1} and $\widetilde{M}^{-1}\widetilde{N}$. The terms M^{-1} and \widetilde{M}^{-1} can be unbounded and we must ensure that the domains and ranges are properly aligned so that the products are defined. Although this is not an insurmountable difficulty, we prefer to work with graphs of unstable systems and their representations in this paper.

DEFINITION 2.2. A plant G has a *right representation* $\begin{bmatrix} M \\ N \end{bmatrix}$ if M and N are causal, bounded operators such that

$$\mathcal{G}\{G\} = \mathcal{R} \left\{ \begin{bmatrix} M \\ N \end{bmatrix} \right\}.$$

The right representation is a *strong right representation* if it has a causal, bounded left inverse. If the representation has the property that $M^*M + N^*N = I$, then the representation is said to be normalized.

DEFINITION 2.3. A plant G has a *left representation* $\begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix}$ if \widetilde{M} and \widetilde{N} are causal, bounded operators such that

$$\mathcal{G}\{G\} = \mathcal{K} \left\{ \begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix} \right\}.$$

The left representation is a *strong left representation* if it has a causal, bounded right inverse. If the representation has the property that $\widetilde{M}\widetilde{M}^* + \widetilde{N}\widetilde{N}^* = I$, then the representation is said to be normalized.

The main goal of this paper is to show that the existence of strong right and left representations is equivalent to stabilizability for linear, causal, discrete-time systems. The proof of existence will make use of certain results of Arveson on inner/outer factorizations in nest algebras. The key technical step involves the factorization of the *adjoint* of a certain causal, bounded operator, which can be viewed as belonging to a nest algebra of *anticausal* operators. The paper goes on to develop the Youla parametrization for stabilizable systems in the form of strong representations of graphs. Some implications of the Youla parametrization theorem are examined and an example of Feintuch is proven to be not stabilizable. It is also proven that if a time-invariant plant is not stabilizable with a time-invariant compensator, then it is not stabilizable with a time-varying compensator. Finally, the problems of extending our proof to continuous-time plants is examined, a necessary condition for stabilizability is given, and a continuous-time, time-invariant plant example of Shefi is proven to be not stabilizable. The paper is organized as follows. In §3 some background on nest algebras is given; in §4 the main results are proven; in §5 implications of the Youla parametrization are presented; and in §6 the case of continuous-time plants is discussed.

3. Nest algebras and inner/outer factorizations. The study of linear, causal systems is aided by some theorems in nest algebras. Nest algebras view the inputs and outputs of the system as a chain of subspaces. The only requirement is that the subspaces be linearly ordered (nested). For example, we can write $\ell_2^m[0, \infty) = \cup_{k=0}^{\infty} \mathcal{M}_k$, where $\mathcal{M}_k = P_m(k)\ell_2^m$ and $\mathcal{M}_0 \supset \mathcal{M}_1 \cdots \supset \mathcal{M}_k \supset \mathcal{M}_{k+1} \cdots \supset 0$. Each of these subspaces has an associated orthogonal projection operator $\Pi_{\mathcal{M}_k}$ and the set of projection operators determines an algebra of operators such that $[I - \Pi_{\mathcal{M}_k}]G\Pi_{\mathcal{M}_k} = 0$ for all $k \in \mathbb{Z}$. In our language, this nest algebra is the set of all bounded, causal operators $G : \ell_2^m \rightarrow \ell_2^m$.

A different choice of subspaces yields a different set of operators in the nest algebra. As a second example consider, $\ell_2^m[0, \infty) = \cup_{k=0}^{\infty} \mathcal{M}_k^*$, where $\mathcal{M}_k^* = [I - P_m(k)]\ell_2^m$ and $0 \subset \mathcal{M}_0^* \subset \mathcal{M}_1^* \cdots \subset \mathcal{M}_k^* \subset \mathcal{M}_{k+1}^* \cdots$. Each of these subspaces has an associated orthogonal projection operator $\Pi_{\mathcal{M}_k^*}$ and the set of projection operators determines an algebra of operators such that $[I - \Pi_{\mathcal{M}_k^*}]G^*\Pi_{\mathcal{M}_k^*} = 0$ for all $k \in \mathbb{Z}$. Note that this nest algebra is the set of all bounded, anticausal operators $G^* : \ell_2^m \rightarrow \ell_2^m$. Also, the adjoint of any operator in the first nest algebra is an operator in the second nest algebra.

Similar constructions are possible for the continuous-time case of $L_2^m[0, \infty)$. In this case, the subspaces are indexed by the real numbers and the nest is called a continuous nest.

Arveson in [1] defined inner/outer factorizations for time-varying operators in a nest algebra. The results of Arveson and others are brought together in a unified fashion in [2]. Few of the properties that define inner and outer are used in this paper. However, the existence of inner/outer factorizations with certain properties is crucial to our proofs. For completeness, we include the following definitions. An operator A is *outer* if the range projection $\Pi_{\overline{\mathcal{R}\{A\}}}$ commutes with the subspace projection $\Pi_{\mathcal{M}_k}$ for all k and $A\mathcal{M}_k$ is dense in $\mathcal{M}_k \cap \mathcal{R}\{A\}$. An operator U is *inner* if U is a partial isometry and U^*U commutes with $\Pi_{\mathcal{M}_k}$ for all k . The following theorem can be found in [1, Thm. 3.2, 3.3 and Cor. 1] and [2, Thms. 14.20 and 14.21].

THEOREM 3.1. *Let \mathcal{N} be a nest algebra. If every \mathcal{M}_k has an immediate successor, then every operator $G \in \mathcal{N}$ has an inner/outer factorization $G = UA$ such that $U \in \mathcal{N}$ is inner, $A \in \mathcal{N}$ is outer, $\overline{\mathcal{R}\{G\}} = \overline{\mathcal{R}\{U\}}$, $\overline{\mathcal{R}\{A\}} = \mathcal{K}\{U\}^\perp$, and $\mathcal{K}\{A\} = \mathcal{K}\{G\}$.*

In addition if $G^*G = A^*A = B^*B$ with B outer, then there exists a partial isometry $V \in \mathcal{N} \cap \mathcal{N}^*$ such that $V^*V = \Pi_{\overline{\mathcal{R}\{A\}}}$, $VV^* = \Pi_{\overline{\mathcal{R}\{B\}}}$, $VA = B$ and $G = WB$ is another inner/outer factorization with $W = UV^*$ inner and B outer.

Since the subspace \mathcal{M}_k has an immediate successor \mathcal{M}_{k+1} , the inner/outer factorizations exist for discrete-time systems. However, this is not the case for continuous-time systems and will be discussed in a later section.

The two examples of nest algebras and Theorem 3.1 assume the operators are “square”, i.e., have equal number of inputs and outputs. We need a factorization result for “tall” operators with more outputs than inputs. The extension is accomplished by packing the operator with zero operators such that the new operator is square and using Theorem 3.1. We will also need the following result [2, Thm. 14.19].

THEOREM 3.2. *If $\{\mathcal{M}_k\}$ is a well-ordered nest, then every positive operator Q factors as A^*A where A is outer.*

Applying Theorem 3.1 to the packed operator we obtain

$$G = \begin{bmatrix} G_1 & 0 \\ G_2 & 0 \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = UA$$

with U inner and A outer. Since $\mathcal{K}\{G\} = \mathcal{K}\{A\}$, we have

$$G = \begin{bmatrix} G_1 & 0 \\ G_2 & 0 \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ A_{21} & 0 \end{bmatrix} = UA.$$

We also have that

$$G^*G = A^*A = \begin{bmatrix} A_{11}^*A_{11} + A_{21}^*A_{21} & 0 \\ 0 & 0 \end{bmatrix}.$$

Since $A_{11}^*A_{11} + A_{21}^*A_{21}$ is a positive, square operator, there exists an outer square operator B_{11} such that

$$G^*G = B^*B = \begin{bmatrix} B_{11}^* & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} B_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

and furthermore B is outer in the larger nest algebra.

By Theorem 3.1, there exists an inner W such that $G = WB$. Hence,

$$G = \begin{bmatrix} G_1 & 0 \\ G_2 & 0 \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} B_{11} & 0 \\ 0 & 0 \end{bmatrix} = WB.$$

Since $\mathcal{K}\{W\}^\perp = \overline{\mathcal{R}\{B\}}$, then

$$G = \begin{bmatrix} G_1 & 0 \\ G_2 & 0 \end{bmatrix} = \begin{bmatrix} W_{11} & 0 \\ W_{21} & 0 \end{bmatrix} \begin{bmatrix} B_{11} & 0 \\ 0 & 0 \end{bmatrix} = WB.$$

By unpacking the above operators, we obtain the following corollary.

COROLLARY 3.3. *Let $G : \ell_2^m \rightarrow \ell_2^n$ ($n \geq m$) be a bounded causal (respectively, anti-causal) operator. Then there exist bounded causal (respectively, anticausal) operators $U : \ell_2^m \rightarrow \ell_2^n$ and $A : \ell_2^m \rightarrow \ell_2^m$ such that $G = UA$, U is a partial isometry, $\overline{\mathcal{R}\{G\}} = \overline{\mathcal{R}\{U\}}$, $\overline{\mathcal{R}\{A\}} = \mathcal{K}\{U\}^\perp$, and $\mathcal{K}\{A\} = \mathcal{K}\{G\}$.*

4. Existence of strong representations. In this section, we prove the main theorem of the paper, which establishes the existence of strong representations for stabilizable systems.

Let G and F be linear and suppose that the closed-loop system $\{G, F\}$ is stable. Then

$$\begin{bmatrix} I & F \\ G & I \end{bmatrix}^{-1} = \begin{bmatrix} (I - FG)^{-1} & -F(I - GF)^{-1} \\ -G(I - FG)^{-1} & (I - GF)^{-1} \end{bmatrix} =: R,$$

where all four elements of R are bounded. This can be seen by writing out the feedback equations with $u_1 = 0$ and $u_2 = 0$ in turn. For the stable (causal) closed-loop system $\{G, F\}$ define the following two operators:

$$(4) \quad P_1 = \begin{bmatrix} I \\ G \end{bmatrix} \begin{bmatrix} (I - FG)^{-1} & -F(I - GF)^{-1} \end{bmatrix}$$

$$(5) \quad P_2 = \begin{bmatrix} F \\ I \end{bmatrix} \begin{bmatrix} -G(I - FG)^{-1} & (I - GF)^{-1} \end{bmatrix}.$$

Note that they are causal and bounded and satisfy

$$\begin{aligned} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} R &= P_1 - \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \\ &= -P_2 + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

It was pointed out in Foias, Georgiou, and Smith [5] that P_1 is the parallel projection operator onto $\mathcal{G}\{G\}$ along $\mathcal{G}^{-1}\{F\}$. In particular $P_1^2 = P_1$. Also, P_2 is the parallel projection operator onto $\mathcal{G}^{-1}\{F\}$ along $\mathcal{G}\{G\}$ and $P_2^2 = P_2$ with $P_1 + P_2 = I$.

We now present our main theorem, which shows that stabilizability implies the existence of strong representations.

THEOREM 4.1. *Let G and F be causal, discrete-time operators and suppose that the closed-loop system $\{G, F\}$ is stable (and causal). Then there exist bounded, causal operators $M, N, X, Y, \widetilde{M}, \widetilde{N}, \widetilde{X},$ and \widetilde{Y} such that*

$$P_1 = \begin{bmatrix} M \\ N \end{bmatrix} \begin{bmatrix} Y & X \end{bmatrix} \text{ and } P_2 = \begin{bmatrix} -\widetilde{X} \\ \widetilde{Y} \end{bmatrix} \begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix}.$$

For any such factorizations, $\begin{bmatrix} M \\ N \end{bmatrix}$ is a strong right representation of G , $\begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix}$ is a strong left representation of G , and the double Bezout identity

$$\begin{aligned} \begin{bmatrix} Y & X \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \begin{bmatrix} M & -\widetilde{X} \\ N & \widetilde{Y} \end{bmatrix} &= \begin{bmatrix} M & -\widetilde{X} \\ N & \widetilde{Y} \end{bmatrix} \begin{bmatrix} Y & X \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \end{aligned}$$

is satisfied. Further, both right and left representations of G can be taken to be normalized.

Proof. Write

$$P_1 = \begin{bmatrix} I \\ G \end{bmatrix} \begin{bmatrix} A_1 & A_2 \end{bmatrix},$$

where

$$A := \begin{bmatrix} A_1 & A_2 \end{bmatrix} := \begin{bmatrix} (I - FG)^{-1} & -F(I - GF)^{-1} \end{bmatrix}.$$

Since A is bounded and causal, its adjoint A^* is bounded and anti-causal. Applying Corollary 3.3 to A^* we have a factorization

$$A^* = \begin{bmatrix} Y^* \\ X^* \end{bmatrix} V^*,$$

where $\begin{bmatrix} Y^* \\ X^* \end{bmatrix}$ is a partial isometry and $\mathcal{K} \left\{ \begin{bmatrix} Y^* \\ X^* \end{bmatrix} \right\}^\perp = \overline{\mathcal{R}} \{V^*\}$. Taking the adjoint we have

$$A = V \begin{bmatrix} Y & X \end{bmatrix}$$

with V and $\begin{bmatrix} Y & X \end{bmatrix}$ being bounded and causal. We now write

$$P_1 = \begin{bmatrix} V \\ GV \end{bmatrix} \begin{bmatrix} Y & X \end{bmatrix} =: \begin{bmatrix} M \\ N \end{bmatrix} \begin{bmatrix} Y & X \end{bmatrix}.$$

Observe that $\begin{bmatrix} M \\ N \end{bmatrix}$ is causal since both V and G are causal operators. We now wish to show that $\begin{bmatrix} M \\ N \end{bmatrix}$ is bounded. Since the adjoint of a partial isometry is a partial isometry, $\begin{bmatrix} Y & X \end{bmatrix}$ is a partial isometry and

$$I - \begin{bmatrix} Y & X \end{bmatrix} \begin{bmatrix} Y^* \\ X^* \end{bmatrix}$$

is an orthogonal projection onto

$$\begin{aligned} \overline{\mathcal{R}} \left\{ \begin{bmatrix} Y & X \end{bmatrix} \right\}^\perp &= \mathcal{K} \left\{ \begin{bmatrix} Y^* \\ X^* \end{bmatrix} \right\} \\ &= \overline{\mathcal{R}} \{V^*\}^\perp = \mathcal{K} \{V\} = \mathcal{K} \left\{ \begin{bmatrix} M \\ N \end{bmatrix} \right\}. \end{aligned}$$

Hence,

$$\begin{aligned} &\begin{bmatrix} M \\ N \end{bmatrix} \\ &= \begin{bmatrix} M \\ N \end{bmatrix} \begin{bmatrix} Y & X \end{bmatrix} \begin{bmatrix} Y^* \\ X^* \end{bmatrix} + \begin{bmatrix} M \\ N \end{bmatrix} \left\{ I - \begin{bmatrix} Y & X \end{bmatrix} \begin{bmatrix} Y^* \\ X^* \end{bmatrix} \right\} \\ (6) \quad &= P_1 \begin{bmatrix} Y^* \\ X^* \end{bmatrix} \end{aligned}$$

and so $\begin{bmatrix} M \\ N \end{bmatrix}$ is bounded because it is the product of two bounded operators. We also deduce from (6) and the fact that

$$\mathcal{R} \left\{ \begin{bmatrix} Y^* \\ X^* \end{bmatrix} \right\} \perp \mathcal{K} \left\{ \begin{bmatrix} Y & X \end{bmatrix} \right\}$$

that

$$\mathcal{R} \left\{ \begin{bmatrix} M \\ N \end{bmatrix} \right\} = \mathcal{R} \{P_1\} = \mathcal{G} \{G\}.$$

Thus, $\begin{bmatrix} M \\ N \end{bmatrix}$ is a right representation of G .

Similar reasoning for P_2 yields the factorization

$$P_2 = \begin{bmatrix} -\tilde{X} \\ \tilde{Y} \end{bmatrix} \begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix},$$

where the operators \tilde{M} , \tilde{N} , \tilde{X} , and \tilde{Y} are causal and bounded and $\begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix}$ is a left representation of G .

Now consider *any* bounded, causal M , N , X , Y , \tilde{M} , \tilde{N} , \tilde{X} , and \tilde{Y} such that

$$P_1 = \begin{bmatrix} M \\ N \end{bmatrix} \begin{bmatrix} Y & X \end{bmatrix}$$

and

$$P_2 = \begin{bmatrix} -\tilde{X} \\ \tilde{Y} \end{bmatrix} \begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix}.$$

Then

$$\begin{aligned} & \begin{bmatrix} M & -\tilde{X} \\ N & \tilde{Y} \end{bmatrix} \begin{bmatrix} Y & X \\ -\tilde{N} & \tilde{M} \end{bmatrix} \\ &= \begin{bmatrix} M \\ N \end{bmatrix} \begin{bmatrix} Y & X \end{bmatrix} + \begin{bmatrix} -\tilde{X} \\ \tilde{Y} \end{bmatrix} \begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix} \\ &= P_1 + P_2 = I. \end{aligned}$$

Both the matrix operators on the left-hand side of the above equation are causal, bounded operators on ℓ_2^{m+n} . They therefore have lower triangular matrix representations. Since their product is equal to the identity, the diagonal blocks must all be nonsingular. Hence, neither operator has a kernel and so they are inverses of each other. Thus,

$$\begin{bmatrix} Y & X \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{X} \\ N & \tilde{Y} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

and the double Bezout identity is satisfied. Since $\begin{bmatrix} -\tilde{X} \\ \tilde{Y} \end{bmatrix}$ is left invertible, it has no kernel and

$$\mathcal{G}\{G\} = \mathcal{K}\{P_2\} = \mathcal{K}\left\{\begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix}\right\}.$$

Hence, $\begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix}$ is a strong left representation of G . Similarly, since $\begin{bmatrix} Y & X \end{bmatrix}$ is right invertible,

$$\mathcal{G}\{G\} = \mathcal{R}\{P_1\} = \mathcal{R}\left\{\begin{bmatrix} M \\ N \end{bmatrix}\right\}$$

and so $\begin{bmatrix} M \\ N \end{bmatrix}$ is a strong right representation of G .

Finally, note that, for *any* strong right representation of G , we can inner/outer factorize

$$\begin{bmatrix} M \\ N \end{bmatrix} = \begin{bmatrix} M_1 \\ N_1 \end{bmatrix} U.$$

Moreover, $\begin{bmatrix} M_1 \\ N_1 \end{bmatrix}$ is also left invertible and has range $\mathcal{G}\{G\}$. Since $\begin{bmatrix} M_1 \\ N_1 \end{bmatrix}$ is a partial isometry it must be a *normalized* strong representation of G . Similarly, a *normalized* left representation can be obtained from an arbitrary strong representation by means of an inner/outer factorization on $\begin{bmatrix} -\tilde{N}^* \\ \tilde{M}^* \end{bmatrix}$. \square

In [15], in the time-invariant case, reduction of fractional representations to co-prime representations was achieved by *two* inner/outer factorizations (one on the H_∞ matrix $\begin{bmatrix} M \\ N \end{bmatrix}$ and the other on the transpose of this matrix). It is easy to check that both these factorizations are necessary in general. It is interesting to note that a similar reduction was achieved in the proof of the above theorem by means of just *one* inner/outer factorization of the adjoint operator in the star algebra.

THEOREM 4.2 (Youla parametrization). *Let G be a discrete-time, causal, possibly time-varying plant G that is stabilizable (i.e., $\{G, F\}$ is stable and causal for some F). Consider any bounded, causal operators $M, N, X, Y, \tilde{M}, \tilde{N}, \tilde{X}$, and \tilde{Y} with $\begin{bmatrix} M \\ N \end{bmatrix}$ a strong right representation of G and $\begin{bmatrix} -\tilde{N} & \tilde{M} \end{bmatrix}$ a strong left representation of G such that the following double Bezout identity holds*

$$\begin{bmatrix} Y & X \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{X} \\ N & \tilde{Y} \end{bmatrix} = \begin{bmatrix} M & -\tilde{X} \\ N & \tilde{Y} \end{bmatrix} \begin{bmatrix} Y & X \\ -\tilde{N} & \tilde{M} \end{bmatrix} \\ = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Then, a compensator F stabilizes G (i.e., $\{G, F\}$ is stable and causal for F) if and only if F has a strong right representation

$$\begin{bmatrix} \tilde{Y} - NQ \\ -\tilde{X} - MQ \end{bmatrix}$$

and a strong left representation

$$\begin{bmatrix} X + Q\tilde{M} & Y - Q\tilde{N} \end{bmatrix}$$

for some causal, bounded Q .

Proof. Assume that F stabilizes G . Then F also has a strong left and right representation that satisfy the double Bezout identity. Thus, there exist bounded, causal operators such that

$$\mathcal{G}\{F\} = \mathcal{R}\left\{\begin{bmatrix} M_F \\ N_F \end{bmatrix}\right\} = \mathcal{K}\left\{\begin{bmatrix} -\tilde{N}_F & \tilde{M}_F \end{bmatrix}\right\}$$

and

$$\begin{bmatrix} Y_F & X_F \\ -\tilde{N}_F & \tilde{M}_F \end{bmatrix} \begin{bmatrix} M_F & -\tilde{X}_F \\ N_F & \tilde{Y}_F \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Define

$$H := \begin{bmatrix} \widetilde{M}_F & -\widetilde{N}_F \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix}.$$

Since the left representation of the compensator is right invertible and the right representation of the plant is left invertible, $\mathcal{R}\{[\widetilde{M}_F \ -\widetilde{N}_F]\} = \ell_2^m$, $\mathcal{K}\{[\begin{smallmatrix} M \\ N \end{smallmatrix}]\} = 0$, $\mathcal{R}\{[\begin{smallmatrix} M \\ N \end{smallmatrix}]\} = \mathcal{G}\{G\}$, and $\mathcal{K}\{[\widetilde{M}_F \ -\widetilde{N}_F]\} = \mathcal{G}^{-1}\{F\}$. Because $\{G, F\}$ is stable, we must have $\mathcal{G}\{G\} + \mathcal{G}^{-1}\{F\} = \ell_2^{m+n}$ and $\mathcal{G}\{G\} \cap \mathcal{G}^{-1}\{F\} = 0$. Therefore, $\mathcal{R}\{H\} = \ell_2^m$ and $\mathcal{K}\{H\} = 0$, which means that H has an inverse H^{-1} . We will now establish that H^{-1} is bounded and causal. Define

$$\Phi := \begin{bmatrix} M \\ N \end{bmatrix} H^{-1} \begin{bmatrix} \widetilde{M}_F & -\widetilde{N}_F \end{bmatrix}.$$

Therefore, $\mathcal{K}\{\Phi\} = \mathcal{G}^{-1}\{F\}$, $\mathcal{R}\{\Phi\} = \mathcal{G}\{G\}$, and $\Phi^2 = \Phi$. Thus, Φ is a parallel projection onto $\mathcal{G}\{G\}$ along $\mathcal{G}^{-1}\{F\}$ (which is the bounded, causal operator P_1 defined in terms of the closed-loop operators in (4)). We now write

$$\begin{aligned} \begin{bmatrix} Y & X \end{bmatrix} \Phi \begin{bmatrix} \widetilde{Y}_F \\ -\widetilde{X}_F \end{bmatrix} &= \\ \begin{bmatrix} Y & X \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} H^{-1} \begin{bmatrix} \widetilde{M}_F & -\widetilde{N}_F \end{bmatrix} \begin{bmatrix} \widetilde{Y}_F \\ -\widetilde{X}_F \end{bmatrix} &= H^{-1}, \end{aligned}$$

where the last equality follows from the double Bezout identities. Then H^{-1} is bounded and causal since it is the product of three bounded, causal operators. Hence,

$$\mathcal{G}\{F\} = \mathcal{K}\left\{\begin{bmatrix} -H^{-1}\widetilde{N}_F & H^{-1}\widetilde{M}_F \end{bmatrix}\right\},$$

$$\begin{bmatrix} Y_F & X_F \\ -H^{-1}\widetilde{N}_F & H^{-1}\widetilde{M}_F \end{bmatrix} \begin{bmatrix} M_F & -\widetilde{X}_F H \\ N_F & \widetilde{Y}_F H \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

and

$$\begin{bmatrix} H^{-1}\widetilde{M}_F & -H^{-1}\widetilde{N}_F \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} = I.$$

Thus, *without loss of generality*, we can assume

$$\begin{bmatrix} \widetilde{M}_F & -\widetilde{N}_F \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} = I.$$

Since we also have

$$\begin{bmatrix} Y & X \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} = I,$$

then

$$\begin{bmatrix} \widetilde{M}_F - Y & -\widetilde{N}_F - X \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} = 0.$$

Define the causal, bounded operator

$$Q := \begin{bmatrix} \widetilde{M}_F - Y & -\widetilde{N}_F - X \end{bmatrix} \begin{bmatrix} -\widetilde{X} \\ \widetilde{Y} \end{bmatrix}.$$

Then

$$\begin{aligned} \begin{bmatrix} Y - Q\widetilde{N} & X + Q\widetilde{M} \end{bmatrix} &= \begin{bmatrix} Y & X \end{bmatrix} + Q \begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix} \\ &= \begin{bmatrix} Y & X \end{bmatrix} + \begin{bmatrix} \widetilde{M}_F - Y & -\widetilde{N}_F - X \end{bmatrix} \begin{bmatrix} -\widetilde{X} \\ \widetilde{Y} \end{bmatrix} \begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix} \\ &= \begin{bmatrix} Y & X \end{bmatrix} \\ &\quad + \begin{bmatrix} \widetilde{M}_F - Y & -\widetilde{N}_F - X \end{bmatrix} \left\{ \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} M \\ N \end{bmatrix} \begin{bmatrix} Y & X \end{bmatrix} \right\} \\ &= \begin{bmatrix} \widetilde{M}_F & -\widetilde{N}_F \end{bmatrix} - \left\{ \begin{bmatrix} \widetilde{M}_F - Y & -\widetilde{N}_F - X \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} \right\} \begin{bmatrix} Y & X \end{bmatrix} \\ &= \begin{bmatrix} \widetilde{M}_F & -\widetilde{N}_F \end{bmatrix}. \end{aligned}$$

Since $\begin{bmatrix} X + Q\widetilde{M} & Y - Q\widetilde{N} \end{bmatrix} = \begin{bmatrix} -\widetilde{N}_F & \widetilde{M}_F \end{bmatrix}$, all stabilizing compensators have a strong *left* representation in the form of the Youla parametrization.

Conversely, choose a bounded, causal operator for Q . This yields a compensator F with a *right* representation

$$\mathcal{G}\{F\} = \mathcal{R} \left\{ \begin{bmatrix} \widetilde{Y} - NQ \\ -\widetilde{X} - MQ \end{bmatrix} \right\}.$$

Choose $e_1 \in \mathcal{D}\{G\}$ and $e_2 \in \mathcal{D}\{F\}$ and calculate the closed-loop system inputs as follows:

$$\begin{aligned} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \begin{bmatrix} I \\ G \end{bmatrix} e_1 + \begin{bmatrix} F \\ I \end{bmatrix} e_2 = \begin{bmatrix} M \\ N \end{bmatrix} w_1 + \begin{bmatrix} -\widetilde{X} - MQ \\ \widetilde{Y} - NQ \end{bmatrix} w_2 \\ &= \begin{bmatrix} M & -\widetilde{X} \\ N & \widetilde{Y} \end{bmatrix} \begin{bmatrix} I & -Q \\ 0 & I \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}. \end{aligned}$$

Because the range of the right representation is the graph of the operator, we are guaranteed that w_1 and w_2 exist.

By direct multiplication of the operators and simplification using the double Bezout identities, we obtain

$$\begin{aligned} &\left\{ \begin{bmatrix} I & Q \\ 0 & I \end{bmatrix} \begin{bmatrix} Y & X \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \right\} \left\{ \begin{bmatrix} M & -\widetilde{X} \\ N & \widetilde{Y} \end{bmatrix} \begin{bmatrix} I & -Q \\ 0 & I \end{bmatrix} \right\} \\ &= \left\{ \begin{bmatrix} M & -\widetilde{X} \\ N & \widetilde{Y} \end{bmatrix} \begin{bmatrix} I & -Q \\ 0 & I \end{bmatrix} \right\} \left\{ \begin{bmatrix} I & Q \\ 0 & I \end{bmatrix} \begin{bmatrix} Y & X \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \right\} \\ &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \end{aligned}$$

Therefore,

$$\begin{bmatrix} M & -\tilde{X} \\ N & \tilde{Y} \end{bmatrix} \begin{bmatrix} I & -Q \\ 0 & I \end{bmatrix}$$

is a bounded, causal operator with a bounded, causal inverse. Hence,

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{bmatrix} I & Q \\ 0 & I \end{bmatrix} \begin{bmatrix} Y & X \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

and

$$\begin{aligned} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} &= \begin{bmatrix} M & 0 \\ 0 & \tilde{Y} - NQ \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \\ &= \begin{bmatrix} M & 0 \\ 0 & \tilde{Y} - NQ \end{bmatrix} \begin{bmatrix} I & Q \\ 0 & I \end{bmatrix} \begin{bmatrix} Y & X \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\ &= \begin{bmatrix} I & F \\ G & I \end{bmatrix}^{-1} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \end{aligned}$$

Thus, $\begin{bmatrix} I & F \\ G & I \end{bmatrix}^{-1}$ is bounded for all bounded Q and the system is closed-loop stable. Hence, all strong *right* representations from the Youla parametrization stabilize the plant.

To complete the proof, we will show that, for the same Q , the strong right and strong left representations correspond to the same operator F . Select an arbitrary, causal, bounded Q . Then the controller defined by

$$\mathcal{G}^{-1}\{F\} = \mathcal{R} \left\{ \begin{bmatrix} -\tilde{X} - MQ \\ \tilde{Y} - NQ \end{bmatrix} \right\}$$

stabilizes G . Furthermore,

$$\begin{bmatrix} M & -\tilde{X} - MQ \\ N & \tilde{Y} - NQ \end{bmatrix}$$

is invertible and has range ℓ_2^{m+n} . Hence,

$$\begin{aligned} \begin{bmatrix} Y - Q\tilde{N} & X + Q\tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{X} - MQ \\ N & \tilde{Y} - NQ \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \iff \begin{bmatrix} I & 0 \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

if and only if $w_1 = 0$. Thus

$$\mathcal{K} \left\{ \begin{bmatrix} Y - Q\tilde{N} & X + Q\tilde{M} \end{bmatrix} \right\} = \mathcal{R} \left\{ \begin{bmatrix} -\tilde{X} - MQ \\ \tilde{Y} - NQ \end{bmatrix} \right\}.$$

This establishes the required equality and the proof is complete. \square

5. Implications of the Youla parametrization. In this section, we will prove two theorems that are implied by the Youla parametrization theorem and show that an example given by Feintuch in [4] is not stabilizable.

First, we will prove the following lemma.

LEMMA 5.1. *If a square operator J is bounded and causal with a bounded inverse K , then K is also causal.*

Proof. Because J and K are bounded, each has a matrix representation. Because J is causal, its matrix representation is lower triangular. Hence,

$$J = \begin{bmatrix} J_{00} & 0 & 0 & \cdots \\ J_{10} & J_{11} & 0 & \cdots \\ J_{20} & J_{21} & J_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

and

$$J^{-1} = K = \begin{bmatrix} K_{00} & K_{01} & K_{02} & \cdots \\ K_{10} & K_{11} & K_{12} & \cdots \\ K_{20} & K_{21} & K_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Because $JK = I$ and all the diagonal blocks are square, we must have $J_{00}K_{00} = I$ and $J_{00}K_{0i} = 0$ for $i > 0$. Thus, $K_{0i} = 0$ for $i > 0$ and we proceed inductively for each row of K to conclude that K is lower triangular and thus, causal. \square

We now present a theorem that can be used to prove that a plant is not stabilizable.

THEOREM 5.2. *If a plant G is stabilizable, then any right (respectively, left) representation that has no kernel (respectively, has full range) is a strong right (respectively, strong left) representation. Furthermore, two strong right (respectively, strong left) representations are related to one another by multiplication on the right (respectively, left) by a bounded, causal, square, invertible operator.*

Proof. Define the right representation of the plant as $\begin{bmatrix} M \\ N \end{bmatrix}$ and the left representation of the plant as $\begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix}$. Since the plant is stabilizable, there exists a strong right representation, $\begin{bmatrix} M_1 \\ N_1 \end{bmatrix}$, and a strong left representation, $\begin{bmatrix} -\widetilde{N}_1 & \widetilde{M}_1 \end{bmatrix}$ such that the following double Bezout identity holds:

$$\begin{bmatrix} Y_1 & X_1 \\ -\widetilde{N}_1 & \widetilde{M}_1 \end{bmatrix} \begin{bmatrix} M_1 & -\widetilde{X}_1 \\ N_1 & \widetilde{Y}_1 \end{bmatrix} = \begin{bmatrix} M_1 & -\widetilde{X}_1 \\ N_1 & \widetilde{Y}_1 \end{bmatrix} \begin{bmatrix} Y_1 & X_1 \\ -\widetilde{N}_1 & \widetilde{M}_1 \end{bmatrix} \\ = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Choose a stabilizing compensator from the Youla parametrization by setting $Q = 0$. Since $\{G, F\}$ is stable, we must have $\mathcal{G}\{G\} + \mathcal{G}^{-1}\{F\} = \ell_2^{m+n}$ and $\mathcal{G}\{G\} \cap \mathcal{G}^{-1}\{F\} = 0$. We also have that $\begin{bmatrix} Y_1 & X_1 \end{bmatrix}$ is right invertible and thus, has full range. Therefore,

$$\mathcal{R} \left\{ \begin{bmatrix} M & -\widetilde{X}_1 \\ N & \widetilde{Y}_1 \end{bmatrix} \right\} = \ell_2^{m+n},$$

$$\mathcal{K} \left\{ \begin{bmatrix} M & -\widetilde{X}_1 \\ N & \widetilde{Y}_1 \end{bmatrix} \right\} = 0,$$

$$\mathcal{K} \left\{ \begin{bmatrix} Y_1 & X_1 \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \right\} = 0,$$

and we can write

$$\begin{aligned} \begin{bmatrix} Y_1 & X_1 \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \begin{bmatrix} M & -\widetilde{X}_1 \\ N & \widetilde{Y}_1 \end{bmatrix} &= \begin{bmatrix} Y_1 M + X_1 N & 0 \\ 0 & \widetilde{N} \widetilde{X}_1 + \widetilde{M} \widetilde{Y}_1 \end{bmatrix} \\ &=: \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix} =: J. \end{aligned}$$

Because $\begin{bmatrix} M & -\widetilde{X}_1 \\ N & \widetilde{Y}_1 \end{bmatrix}$ has full range, we have

$$\begin{aligned} \mathcal{R}\{J_1\} &= \mathcal{R}\{Y_1 M + X_1 N\} = \mathcal{R} \left\{ \begin{bmatrix} Y_1 & X_1 \end{bmatrix} \begin{bmatrix} M & -\widetilde{X}_1 \\ N & \widetilde{Y}_1 \end{bmatrix} \right\} \\ &= \mathcal{R} \left\{ \begin{bmatrix} Y_1 & X_1 \end{bmatrix} \right\} = \ell_2^m. \end{aligned}$$

Similarly $\mathcal{R}\{J_2\} = \ell_2^n$, and so $\mathcal{R}\{J\} = \ell_2^{m+n}$. We also have that $\mathcal{K}\{J\} = 0$, so J is one-to-one and onto. A consequence of the open mapping theorem is [14, Thm. 5.10], a bounded operator that is one-to-one and maps a Banach space onto a Banach space has a bounded inverse. Therefore, J^{-1} exists and is bounded. By the lemma, J^{-1} is also causal as are J_1^{-1} and J_2^{-1} .

The first part of the theorem now follows by noting that

$$\begin{bmatrix} J_1^{-1} Y_1 & J_1^{-1} X_1 \\ -\widetilde{N} & \widetilde{M} \end{bmatrix} \begin{bmatrix} M & -\widetilde{X}_1 J_2^{-1} \\ N & \widetilde{Y}_1 J_2^{-1} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Thus $\begin{bmatrix} M \\ N \end{bmatrix}$ and $\begin{bmatrix} -\widetilde{N} & \widetilde{M} \end{bmatrix}$ have left and right inverses, respectively.

To prove the second part of the theorem, we note that

$$\begin{bmatrix} Y_1 & X_1 \\ -J_2^{-1} \widetilde{N} & J_2^{-1} \widetilde{M} \end{bmatrix} \begin{bmatrix} M J_1^{-1} & -\widetilde{X}_1 \\ N J_1^{-1} & \widetilde{Y}_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

and define

$$\begin{aligned} \Psi &:= \begin{bmatrix} M J_1^{-1} & -\widetilde{X}_1 \\ N J_1^{-1} & \widetilde{Y}_1 \end{bmatrix} \begin{bmatrix} Y_1 & X_1 \\ -J_2^{-1} \widetilde{N} & J_2^{-1} \widetilde{M} \end{bmatrix} \\ &= \begin{bmatrix} M & -\widetilde{X}_1 \\ N & \widetilde{Y}_1 \end{bmatrix} \begin{bmatrix} J_1^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & J_2^{-1} \end{bmatrix} \begin{bmatrix} Y_1 & X_1 \\ -\widetilde{N} & \widetilde{M} \end{bmatrix}. \end{aligned}$$

Since $\mathcal{K}\{\Psi\} = 0$ and $\Psi^2 = \Psi$, then $\Psi = I$. By direct computation we have

$$\begin{bmatrix} Y_1 & X_1 \\ -J_2^{-1} \widetilde{N} & J_2^{-1} \widetilde{M} \end{bmatrix} \begin{bmatrix} M_1 & -\widetilde{X}_1 \\ N_1 & \widetilde{Y}_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Since inverses are unique, we have

$$\begin{bmatrix} -J_2^{-1}\tilde{N} & J_2^{-1}\tilde{M} \end{bmatrix} = \begin{bmatrix} -\tilde{N}_1 & \tilde{M}_1 \end{bmatrix} \text{ and } \begin{bmatrix} MJ_1^{-1} \\ NJ_1^{-1} \end{bmatrix} = \begin{bmatrix} M_1 \\ N_1 \end{bmatrix}. \quad \square$$

We now apply the above result to show that an example presented by Feintuch in [4] is not stabilizable. Consider the following single-input, single-output, unbounded operator

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ & \vdots & & & \ddots \end{bmatrix}.$$

Let

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1/5 \\ & \vdots & & & \ddots \end{bmatrix}$$

and

$$N = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ & \vdots & & & \ddots \end{bmatrix}.$$

It can be seen that $\begin{bmatrix} M \\ N \end{bmatrix}$ is a right representation of G . In addition, $\mathcal{K}\{\begin{bmatrix} M \\ N \end{bmatrix}\} = 0$. Following Feintuch[4], assume the existence of a bounded, causal left inverse $\begin{bmatrix} Y & X \end{bmatrix}$. Since N has zeros on its diagonal, XN will also have zeros on its diagonal. Therefore, YM must have ones on its diagonal. This implies that the diagonal of Y is $\{1, 2, 3, 4, \dots\}$, which implies that Y is unbounded. Thus, $\begin{bmatrix} M \\ N \end{bmatrix}$ has no left inverse. By Theorem 5.2, if G is stabilizable, then $\begin{bmatrix} M \\ N \end{bmatrix}$ must have a left inverse. Thus, G is not stabilizable.

Finally, we prove the following theorem about time-invariant systems.

THEOREM 5.3. *Let G be time-invariant and causal. If G is not stabilizable with a time-invariant, causal compensator F , then it is not stabilizable with a time-varying, causal F .*

Proof. In this proof we will be changing our viewpoint between time and frequency domain several times. Thus, we present the following notation. Let M be a shift-invariant operator in the time domain and $\mathbf{M}(z)$ be the corresponding H_∞ matrix multiplication operator. Also, let $y(t)$ be a signal of finite energy in the time domain and $\mathbf{y}(z)$ be the corresponding H_2 vector function.

In the first part of the proof, we construct a sequence of inputs to show that the right representation is not left invertible. The argument used is closely related to a proof in [7, Prop. 7].

From the Beurling-Lax Theorem, we can write (in the frequency domain)

$$\mathcal{G}\{G(z)\} = \begin{bmatrix} M(z) \\ N(z) \end{bmatrix} H_2^m,$$

where $M(z)$ and $N(z)$ are matrices over H_∞ such that

$$M(z)^* M(z) + N(z)^* N(z) = I$$

and H_p^m is the Hardy p -space of vector valued functions on the disc. Since G is not stabilizable by a time-invariant, causal F , $\begin{bmatrix} M(z) \\ N(z) \end{bmatrix}$ is not left invertible as a matrix over H_∞ . This means that $\inf_{|z|<1} \sigma_{\min} \begin{bmatrix} M(z) \\ N(z) \end{bmatrix} = 0$ by the matrix valued Corona (Fuhrmann) theorem. Thus, we can find a sequence z_i with $|z_i| < 1$ and complex m -vectors x_i of unit norm such that $\begin{bmatrix} M(z_i) \\ N(z_i) \end{bmatrix} x_i \rightarrow 0$ as $i \rightarrow \infty$. Construct the $(H_2^m)^\perp = L_2^m \ominus H_2^m$ vectors

$$y_i''(z) = \left(\frac{c_i}{z - z_i} \right) x_i,$$

where c_i are complex constants chosen so that $\|y_i''(z)\|_2 = 1$. Then

$$\begin{aligned} w_i''(z) &:= \begin{bmatrix} M(z) \\ N(z) \end{bmatrix} y_i''(z) \\ &= \begin{bmatrix} M(z_i) \\ N(z_i) \end{bmatrix} y_i''(z) + \begin{bmatrix} M(z) - M(z_i) \\ N(z) - N(z_i) \end{bmatrix} y_i''(z) \\ &=: (w_i'')^-(z) + (w_i'')^+(z), \end{aligned}$$

where $\|w_i''(z)\|_2 = 1$, $(w_i'')^-(z) \in (H_2^m)^\perp$, and $(w_i'')^+(z) \in H_2^m$. Furthermore, $\|(w_i'')^-(t)\|_2 \rightarrow 0$ as $i \rightarrow \infty$.

Consider the $\ell_2^m(-\infty, -1)$ vector corresponding to $y_i''(z)$. There exist normalized truncations $y_i'(t) \in \ell_2^m(-k_i, -1)$ such that $\|y_i'(t)\|_2 = 1$ and

$$w_i'(z) := \begin{bmatrix} M(z) \\ N(z) \end{bmatrix} y_i'(z) =: (w_i')^+(z) + (w_i')^-(z),$$

where $(w_i')^-(z) \in (H_2^m)^\perp$, $(w_i')^+(z) \in H_2^m$, and $\|(w_i')^-(z)\|_2 \rightarrow 0$. This follows since $\begin{bmatrix} M(z) \\ N(z) \end{bmatrix}$ is a bounded operator on L_2^m of the unit circle.

Shifting $y_i'(t)$ yields $y_i(t) \in \ell_2^m(0, k_i - 1)$ with $\|y_i(t)\|_2 = 1$ and

$$w_i(t) := \begin{bmatrix} M \\ N \end{bmatrix} y_i(t) =: w_i^-(t) + w_i^+(t),$$

where $w_i^-(t) \in \ell_2^m(0, k_i - 1)$, $w_i^+(t) \in \ell_2^m(k_i, \infty)$, and $\|w_i^-(t)\|_2 \rightarrow 0$.

The proof now proceeds by contradiction. Suppose G is stabilizable with a time-varying, causal F . Then $\begin{bmatrix} M \\ N \end{bmatrix}$ is a strong right representation of G by Theorem 5.2. Hence, there exists bounded, causal operators Y and X such that

$$\begin{bmatrix} Y & X \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} = I.$$

Let $\| \begin{bmatrix} Y & X \end{bmatrix} \| = c$. Then

$$\begin{aligned} 1 &= \|y_i(t)\|_2 = \| [I - P_m(k_i)] y_i(t) \|_2 \\ &= \left\| [I - P_m(k_i)] \begin{bmatrix} Y & X \end{bmatrix} \begin{bmatrix} M \\ N \end{bmatrix} y_i(t) \right\|_2 \\ &= \left\| [I - P_m(k_i)] \begin{bmatrix} Y & X \end{bmatrix} [I - P_{m+n}(k_i)] \begin{bmatrix} M \\ N \end{bmatrix} y_i(t) \right\|_2 \\ &\leq c \left\| [I - P_{m+n}(k_i)] \begin{bmatrix} M \\ N \end{bmatrix} y_i(t) \right\|_2 = c \|w_i^-(t)\|_2 \rightarrow 0, \end{aligned}$$

which is a contradiction. \square

6. Continuous-time plants. In this section, we will discuss problems encountered in extending our proof to continuous-time plants, give a necessary condition for a linear plant to be stabilizable with a linear compensator, and prove that a continuous-time plant of Shefi is not stabilizable.

Our proof of the equivalence of stabilizability and the existence of strong right and strong left system representations fails for continuous-time plants in two important steps. The first step where the proof fails is that inner/outer factorizations do not exist for continuous-time operators as stated in the following theorem found in [2, Thm. 14.20].

THEOREM 6.1. *Let \mathcal{N} be a nest algebra. If there exists \mathcal{M}_k that has no immediate successor, then there exists an operator $G \in \mathcal{N}$ that does not have an inner/outer factorization.*

For a continuous-time nest algebra, the subspaces have no immediate successor and the inner/outer factorizations may not exist. Thus, our proof cannot be used in the continuous-time case. However, the properties of U that we use in our proof are that U is a partial isometry with $\overline{\mathcal{R}}\{G\} = \overline{\mathcal{R}}\{U\}$ and that $\overline{\mathcal{R}}\{A\} = \mathcal{K}\{U\}^\perp$ and $\mathcal{K}\{A\} = \mathcal{K}\{G\}$. Thus, if a factorization can be found that possesses these properties, this step in the proof would carry over to the continuous-time case.

The second step where our proof fails is in proving

$$\begin{bmatrix} Y & X \\ -\tilde{N} & \tilde{M} \end{bmatrix} \begin{bmatrix} M & -\tilde{X} \\ N & \tilde{Y} \end{bmatrix} = I.$$

In this step, we used the fact that all bounded, causal operators on ℓ_2^m have a lower triangular matrix representation.

The following is a necessary condition for stabilizability in both the continuous-time and discrete-time cases.

THEOREM 6.2. *If a continuous-time or discrete-time plant G is linear, possibly time-varying, and stabilizable, then it has a closed graph.*

Proof. Since G is stabilizable, choose a compensator F such that the closed-loop system $\{G, F\}$ is stable. From (4), the parallel projection P_1 onto $\mathcal{G}\{G\}$ along $\mathcal{G}^{-1}\{F\}$ is defined in terms of the closed-loop operators. Thus,

$$\begin{bmatrix} e_1 \\ y_1 \end{bmatrix} = P_1 \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

where P_1 is a bounded, linear operator that maps ℓ_2^{m+n} onto $\mathcal{G}\{G\}$ and $P_1^2 = P_1$. If $e_1 \in \mathcal{D}\{G\}$ and $y_1 = Ge_1$ then, $\begin{bmatrix} e_1 \\ y_1 \end{bmatrix} = P_1 \begin{bmatrix} e_1 \\ y_1 \end{bmatrix}$. Now consider a Cauchy sequence

$\begin{bmatrix} e_{1i} \\ y_{1i} \end{bmatrix}$ on $\mathcal{G}\{G\}$ with a limit point $\begin{bmatrix} e'_1 \\ y'_1 \end{bmatrix}$. Since P_1 is a bounded, linear operator, it has a closed graph. Thus,

$$\left(\begin{bmatrix} e_{1i} \\ y_{1i} \\ P_1 \begin{bmatrix} e_{1i} \\ y_{1i} \end{bmatrix} \end{bmatrix} \right) = \begin{pmatrix} e_{1i} \\ y_{1i} \\ e_{1i} \\ y_{1i} \end{pmatrix} \rightarrow \begin{pmatrix} e'_1 \\ y'_1 \\ e'_1 \\ y'_1 \end{pmatrix} \in \mathcal{G}\{P_1\}.$$

Hence,

$$\begin{bmatrix} e'_1 \\ y'_1 \end{bmatrix} = P_1 \begin{bmatrix} e'_1 \\ y'_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} e'_1 \\ y'_1 \end{bmatrix} \in \mathcal{G}\{G\}$$

and the graph of G is closed. \square

Note how the plant inherits the property of having a closed graph from the closed graph of the closed-loop system. This is not true for nonlinear systems since there exist stable, nonlinear plants that do not have a closed graph.

We now present a continuous-time, time-invariant plant example by Shefi from [11, Ex. 1.1–1] and prove that it is not stabilizable with any linear, possibly time-varying compensator. Consider a single-input, single-output, continuous-time plant G defined on all piecewise continuous functions over $[0, \infty)$ having only a finite number of simple jump discontinuities in a finite time. For an input x , the output Gx at any time t is the algebraic sum of the jumps of the input x from 0 up to the present time t . Since we are interested only in L_2 signals, the $\mathcal{D}\{G\}$ is the intersection of L_2 with the above piecewise continuous functions that also yield L_2 output functions.

It is easy to show that the plant G is linear and time-invariant. However, we will show that the graph is not closed. First note that if a pulse of finite amplitude and duration is applied to the input, the output is a pulse of the same amplitude and duration. Since the continuous functions are dense in L_2 and the output to any continuous L_2 function is 0, we can construct a Cauchy sequence of continuous functions on $\mathcal{G}\{G\}$ such that the input sequence converges to a pulse and the output sequence is always 0. Thus, the graph is not closed nor is $\overline{\mathcal{G}}\{G\}$ the graph of any operator because $\overline{\mathcal{G}}\{G\}$ has multiple possible output functions for one input function. Since the graph is not closed, the plant is not stabilizable with any linear, possibly time-varying compensator.

REFERENCES

- [1] W. ARVESON, *Interpolation problems in nest algebras*, J. Functional Anal., 20 (1975), pp. 208–233.
- [2] K. DAVIDSON, *Nest Algebras*, Pitman Research Notes in Mathematics Series, Longman Scientific and Technical, Essex, UK, 1988.
- [3] C. A. DESOER, R. W. LIU, AND R. SAEKS, *Feedback system design: the fractional representation approach to analysis and design*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 399–412.
- [4] A. FEINTUCH, *Coprime factorization of discrete time-varying systems*, System Control Lett., 7 (1986), pp. 49–50.
- [5] C. FOIAS, T. T. GEORGIOU, AND M. C. SMITH, *Geometric techniques for robust stabilization of linear time-varying systems*, in Proceedings of the IEEE Conference on Decision and Control, Honolulu, Hawaii, December 1990, pp. 2868–2873.
- [6] T. T. GEORGIOU AND M. C. SMITH, private communication, 1990.
- [7] ———, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.

- [8] J. HAMMER, *Non-linear systems, stabilization, and coprimeness*, Internat. J. Control, 42 (1985), pp. 1–20.
- [9] ———, *Fraction representations of non-linear systems: a simplified approach*, Internat. J. Control, 46 (1987), pp. 455–472.
- [10] Y. INOUE, *Parametrization of compensators for linear systems with transfer functions of bounded type*, Technical Rept. 88-01, Faculty of Engineering Science, Osaka University, Osaka, Japan, March 1988.
- [11] T. KAILATH, *Linear Systems*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1980.
- [12] K. POOLLA AND P. KHARGONEKAR, *Stabilizability and stable-proper factorizations for linear time-varying systems*, SIAM J. Control Optim., 25 (1987), pp. 723–36.
- [13] M. A. ROTEA AND P. P. KHARGONEKAR, *Stabilizability of linear time-varying and uncertain linear systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 884–887.
- [14] W. RUDIN, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1987.
- [15] M. C. SMITH, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–7.
- [16] M. VERMA, *Coprime fractional representations and stability of non-linear feedback systems*, Internat. J. Control, 48 (1988), pp. 897–918.
- [17] M. VIDYASAGAR, *Control Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
- [18] M. VIDYASAGAR, H. SCHNEIDER, AND B. A. FRANCIS, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 880–94.

AN ALGORITHM FOR A CLASS OF CONTINUOUS LINEAR PROGRAMS*

MALCOLM C. PULLAN†

Abstract. This paper discusses a class of continuous linear programs posed in a function space called separated continuous linear programs (SCLP). A dual linear program and a corresponding discrete approximation are introduced followed by a discussion of their properties. The discrete approximation gives rise to an improvement step which is constructed from any given feasible (nonoptimal) solution for SCLP. A strong duality result follows from this. There are a variety of possible implementations of an algorithm for solving SCLP problems using this improvement step. Finally some computational results are given from one possible implementation.

Key words. continuous linear program, duality, discrete approximation, simplex-like algorithm.

AMS subject classifications. 49A55, 49B36, 49D99, 90C48.

1. Introduction. The following problem,

$$\begin{aligned} \text{CLP: maximize } & \int_0^T c^T(t)x(t) dt \\ \text{subject to } & B(t)x(t) + \int_0^t K(s,t)x(s) ds \leq b(t), \\ & x(t) \geq 0, \quad t \in [0, T], \end{aligned}$$

was first considered by Bellman [7] in 1957 in an attempt to model some economic processes which he called “Bottleneck Processes.” Since then the problem has been studied by a number of authors whose work can be loosely divided into two areas, those concerned with establishing strong duality theorems and those concerned with computational methods.

In the area of duality, Bellman himself established a weak duality result. This was later followed by, for example, Tyndall [19], Levinson [15], and Grinold [12], all of whom gave strong duality results with varying algebraic restrictions on the problem. The dual problems which these authors have considered have all been in the space of bounded measurable functions.

On the computational side, two main approaches can be found. The simpler of the two is to solve the problem via a series of discrete approximations, called discretizations, to the original CLP. Such is the approach taken by Buie and Abrahm [8]. These discrete approximations work by partitioning the interval $[0, T]$ into a finite number of smaller subintervals, normally of equal size. From the point of view of actually solving problems, this approach has had the most success as the solutions for the discretizations converge to the solution for CLP as the discretizations become finer. However, these methods fail to address the infinite-dimensional nature of the problem and this has two major disadvantages. First, the computation times involved can be very large as a result of solving a very large finite-dimensional linear program, and, second, the problem may have an optimal solution of a simple form that is obscured in the discretized version.

* Received by the editors October 23, 1991; accepted for publication (in revised form) April 7, 1992.

† Judge Institute of Management Studies, Mill Lane, Cambridge CB2 1RX, United Kingdom. This work was partially supported by Trinity College, Cambridge.

These observations have lead a number of authors to attempt to solve the problem directly. The natural approach here has been to extend the simplex method for finite-dimensional linear programs to the continuous-time or infinite-dimensional problem. This involves extending such concepts as “basic solutions,” “dual variables,” and “pivots.” This approach was started by Lehman [14] and continued by Drews [10], Hartberger [13], and Segers [18]. Major progress was not made in this direction until the work of Perold [17], later continued by Anstreicher [6]. The algorithm described by Perold, however, is both complicated and incomplete, reflecting the difficult nature of the problem.

In this paper we consider a subclass of CLP, called separated continuous linear programs SCLP. This is defined as follows (we choose to minimize rather than maximize as this is how most optimization problems are now stated),

$$\begin{aligned}
 \text{SCLP: minimize} \quad & \int_0^T c^T(t)x(t) dt \\
 \text{(1)} \quad & \text{subject to} \quad \int_0^t Gx(s) ds + y(t) = a(t), \\
 \text{(2)} \quad & Hx(t) + z(t) = b(t), \\
 & x(t), y(t), z(t) \geq 0, \quad t \in [0, T].
 \end{aligned}$$

Here $x(t)$, $z(t)$, $b(t)$ and $c(t)$ are bounded measurable functions and $y(t)$ and $a(t)$ are continuous functions. The dimensions of $x(t)$, $y(t)$ and $z(t)$ are n_1 , n_2 , and n_3 , respectively. We let $n = n_1 + n_2 + n_3$ and $\omega^T(t) = (x^T(t), y^T(t), z^T(t))$.

This problem was first introduced by Anderson [1] in an attempt to model job-shop scheduling problems. He gave a preliminary combined primal-dual algorithm and a few simple problems were solved; however, full implementation of an algorithm along these lines proved difficult. In 1986 Anderson and Philpott [4] restricted the problem of SCLP even further to the important class of continuous network programs and developed, for the first time, an algorithm for solving these problems directly on a computer, which addresses the infinite dimensionality of the problem. This is made possible by allowing only piecewise linear components of $a(t)$ and $c(t)$ and piecewise constant $b(t)$.

In this paper we develop an algorithm for solving SCLP under the same piecewise linearity constraints as for the network program. Our algorithm combines elements from both the discrete and simplex method approaches outlined above. Our first step in developing the algorithm is to introduce a dual problem, SCLP* (§2), which, in some sense, contains the one considered by Tyndall [19] and others. A new discretization for SCLP is then introduced (§3), which differs from the standard one. This turns out to be a natural discretization for SCLP*. This is followed by a discussion of its properties, which includes a convergence result for a discrete algorithm. We then establish an improvement step by showing how to construct an improved feasible solution starting from any given feasible (non-optimal) solution for SCLP (§4). This leads to a strong duality result.

Many possible implementations of an algorithm can arise from this approach and an outline is given of some of them in §5. One of these is very similar to the algorithm for networks given in [4]. Finally, some computational results are given from solving two problems using a straightforward (although not very efficient) method. One of these examples shows that it is possible to obtain the exact solution in a finite number of steps. This extends the problem class for which direct infinite-dimensional solutions are available from networks to SCLP.

We now introduce some standard definitions and notation which we will use throughout the remainder of this paper.

DEFINITION 1.

1. A set $P = \{t_0, \dots, t_m\}$ is said to be a *partition* of $[0, T]$ if

$$0 = t_0 < t_1 < \dots < t_m = T.$$

2. A partition Q of $[0, T]$ is said to be a *refinement* of a partition P of $[0, T]$ if $P \subseteq Q$.

3. The *norm*, $\|P\|$, of a partition $P = \{t_0, \dots, t_m\}$, is defined by

$$\|P\| = \max_i (t_i - t_{i-1}).$$

4. We say that a function $f(t)$ is piecewise constant (linear) with a partition $P = \{t_0, \dots, t_m\}$, if $f(t)$ is constant (linear) on $[t_{i-1}, t_i)$ for $i = 1, \dots, m$. We say that f is piecewise constant (linear) on $[0, T]$ if f is piecewise constant (linear) with some partition of $[0, T]$.

5. For any linear program LP we will write $V[\text{LP}]$ for its optimal value.

6. For a function $f(t)$ we will use the notations

$$f(t-) = \lim_{s \rightarrow t-} f(s) \quad \text{and} \quad f(t+) = \lim_{s \rightarrow t+} f(s),$$

when the above limits exist.

2. The dual problem. In this section we introduce a dual linear program for SCLP and give a weak duality result (Lemma 2.1). The dual problem SCLP* is defined as follows

$$\begin{aligned} \text{SCLP*}: \quad & \text{maximize} \quad - \int_0^T d\pi^T(t)a(t) - \int_0^T \eta^T(t)b(t) dt \\ & \text{subject to} \quad c(t) - G^T\pi(t) + H^T\eta(t) \geq 0, \\ & \quad \eta(t) \geq 0, \text{ a.e. on } [0, T], \\ & \quad \pi(t) \text{ monotonic increasing and right continuous} \\ & \quad \text{on } [0, T] \text{ with } \pi(T) = 0. \end{aligned}$$

Here the components of $\eta(t)$ are in $L_1[0, T]$. It will be convenient to write $\theta(t)$ for the complete set of variables of SCLP* and $\psi(t)$ for the left-hand side of the first constraint. Thus

$$\theta^T(t) = (\pi^T(t), \eta^T(t)) \text{ and } \psi(t) = c(t) - G^T\pi(t) + H^T\eta(t).$$

LEMMA 2.1 (Weak duality). $V[\text{SCLP*}] \leq V[\text{SCLP}]$.

Proof. Suppose $\omega(t)$ is feasible for SCLP and $\theta(t)$ is feasible for SCLP*. Then

$$\begin{aligned} & - \int_0^T d\pi^T(t)a(t) - \int_0^T \eta^T(t)b(t) dt \\ & = - \int_0^T d\pi^T(t) \left(\int_0^t Gx(s) ds + y(t) \right) - \int_0^T \eta^T(t)(Hx(t) + z(t)) dt \\ & = \int_0^T (G^T\pi(t) - H^T\eta(t))^T x(t) dt - \int_0^T d\pi^T(t)y(t) - \int_0^T \eta^T(t)z(t) dt, \end{aligned}$$

by a standard result (see Dunford and Schwartz [11, p. 154]). Hence

$$\begin{aligned} & \int_0^T c^T(t)x(t) dt + \int_0^T d\pi^T(t)a(t) + \int_0^T \eta^T(t)b(t) dt \\ &= \int_0^T \psi^T(t)x(t) dt + \int_0^T d\pi^T(t)y(t) + \int_0^T \eta^T(t)z(t) dt \\ &\geq 0, \end{aligned}$$

by the feasibility of $\omega(t)$ and $\theta(t)$. This gives the result. \square

Another dual linear program for SCLP is given in Anderson and Nash [3] (for the form of SCLP that is a maximization rather than a minimization) and is defined as follows:

$$\begin{aligned} \text{SCLP}^*: \text{ maximize } & - \int_0^T a^T(t)u(t) dt - \int_0^T b^T(t)v(t) dt \\ \text{subject to } & c(t) + \int_t^T G^T u(s) ds + H^T v(t) \geq 0, \\ & u(t), v(t) \geq 0, \quad t \in [0, T], \end{aligned}$$

with $u(t)$ and $v(t)$ in the space of bounded measurable functions. It is worth noting that if $\theta(t)$ is feasible for SCLP* and $\pi(t)$ is absolutely continuous, then

$$u(t) = \dot{\pi}(t) \quad \text{and} \quad v(t) = \eta(t),$$

is feasible for SCLP*' and the objective function values are the same in their respective linear programs. Conversely, if $u(t)$ and $v(t)$ are feasible for SCLP*' then

$$\pi(t) = - \int_t^T u(s) ds \quad \text{and} \quad \eta(t) = v(t),$$

is feasible for SCLP* and again the two solutions have the same objective function value in their respective linear programs. Because of the above weak duality result, this implies that previous strong duality results between SCLP and SCLP*', such as those given in Tyndall [19], carry over to SCLP and SCLP* under the same restrictions.

Yet another dual linear program with a corresponding strong duality theorem has been considered by Papageorgiou [16]; however, he restricts the variables of the primal to be of bounded variation. Also, his dual problem has all its variables in the space of functions of bounded variation and this makes it possible to construct a feasible solution for this dual linear program from one for SCLP*. The resulting costs of the two solutions are, however, different.

3. Discretizations. In this section we introduce two discretizations for SCLP, the standard and natural one, DP, and a new, less obvious one, AP. This new discretization will turn out to have many interesting properties; for example we will show that its optimal value provides a lower bound for the optimal value of SCLP. However, unlike DP, a solution for AP does not directly give a feasible solution for SCLP with the same cost, but instead one can be constructed in a fairly simple procedure with cost difference that tends to zero with decreasing partition norm. These ideas will lead to the improvement step for SCLP which we describe in the

next section. The proof of the results in §3.2 will rely heavily on the weak duality result for SCLP and SCLP* established in the previous section.

Before proceeding we must make an assumption on the nature of the functions $a(t)$, $b(t)$ and $c(t)$.

ASSUMPTION 1. *The functions $a(t)$ and $c(t)$ are piecewise linear and the function $b(t)$ is piecewise constant on $[0, T]$.*

We shall assume this restriction throughout the rest of this paper. The possibility of weakening or removing this restriction is still an open question. It is shown in Anderson, Nash, and Perold [2] that if the feasible region for SCLP is bounded then Assumption 1 guarantees the existence of an optimal solution for SCLP in which the components of $x(t)$ are piecewise constant. This leads us to introduce the following assumption, which will occasionally be required.

ASSUMPTION 2. *The feasible region for SCLP is bounded.*

In what follows we will consider a fixed arbitrary partition $P = \{t_0, \dots, t_m\}$ of $[0, T]$ so that $a(t)$ and $c(t)$ are piecewise linear and $b(t)$ is piecewise constant with partition P . Given P , we now let

$$Q = \left\{ t_0, \frac{t_0 + t_1}{2}, t_1, \dots, t_{m-1}, \frac{t_{m-1} + t_m}{2}, t_m \right\},$$

i.e., P with each interval split in half.

3.1. The standard discretization. We now state the standard discretization, $DP(P)$, and give some of its properties. This is the discretization used in the early duality work, such as Tyndall [19], and in the computational results of Buie and Abrham [8].

$$\begin{aligned} DP(P): \quad & \text{minimize} \quad \sum_{i=1}^m (t_i - t_{i-1}) \hat{x}^T(t_{i-1}+) c\left(\frac{t_i + t_{i-1}}{2}\right) \\ & \text{subject to} \quad (t_1 - t_0) G \hat{x}(t_0+) + \hat{y}(t_1) = a(t_1), \\ & \quad (t_i - t_{i-1}) G \hat{x}(t_{i-1}+) + \hat{y}(t_i) - \hat{y}(t_{i-1}) = a(t_i) - a(t_{i-1}), \\ & \quad \quad \quad i = 2, \dots, m, \\ & \quad H \hat{x}(t_{i-1}+) + \hat{z}(t_{i-1}+) = b(t_{i-1}+), \quad i = 1, \dots, m, \\ & \quad \hat{x}(t_{i-1}+), \hat{y}(t_i), \hat{z}(t_{i-1}+) \geq 0, \quad i = 1, \dots, m. \end{aligned}$$

For brevity we will write this in matrix form as

$$\begin{aligned} DP(P): \quad & \text{minimize} \quad \hat{c}_D^T \hat{\omega}_D \\ & \text{subject to} \quad A_D \hat{\omega}_D = \hat{b}_D, \\ & \quad \hat{\omega}_D \geq 0, \end{aligned}$$

for some appropriate matrix A_D and vectors \hat{c}_D and \hat{b}_D . The labelling of the variables as $\hat{x}(t_i)$ and $\hat{y}(t_i)$, etc., is for convenience and does not mean that they explicitly refer to a function but rather in an implicit way as shown in the following definition. For this use we will also define $\hat{y}(t_0) = a(t_0)$.

DEFINITION 2. Let $P = \{t_0, \dots, t_m\}$ be a partition of $[0, T]$. Suppose we have a set of $m + 1$ variables called $\hat{f}(t_0), \hat{f}(t_1), \dots, \hat{f}(t_m)$ (e.g., \hat{y} above), then the function $f(t)$ defined by

$$f(t) = \left(\frac{t_i - t}{t_i - t_{i-1}} \right) \hat{f}(t_{i-1}) + \left(\frac{t - t_{i-1}}{t_i - t_{i-1}} \right) \hat{f}(t_i), \text{ for } t \in [t_{i-1}, t_i], \text{ for } i = 1, \dots, m,$$

is called the *piecewise linear extension* of \hat{f} .

Similarly, if we have a set of $2m$ variables called $\hat{f}(t_0+), \hat{f}(t_1-), \hat{f}(t_1+), \dots, \hat{f}(t_m-)$, then the function $f(t)$ defined by

$$f(t) = \begin{cases} \hat{f}(t_i+), & \text{if } t = t_0, \dots, t_{m-1}, \\ 0, & t = T, \\ \left(\frac{t_i - t}{t_i - t_{i-1}}\right) \hat{f}(t_{i-1}+) + \left(\frac{t - t_{i-1}}{t_i - t_{i-1}}\right) \hat{f}(t_i-), & \text{for } t \in (t_{i-1}, t_i), \\ & i = 1, \dots, m, \end{cases}$$

is also called the *piecewise linear extension* of \hat{f} .

Finally, if we have a set of m variables called $\hat{f}(t_0+), \hat{f}(t_1+), \dots, \hat{f}(t_{m-1}+)$ (e.g., \hat{x} or \hat{z} above), then the function $f(t)$ defined by

$$f(t) = \begin{cases} \hat{f}(t_{m-1}+), & t = T, \\ \hat{f}(t_{i-1}+), & t \in [t_{i-1}, t_i), \text{ for } i = 1, \dots, m, \end{cases}$$

is called the *piecewise constant extension* of \hat{f} .

It is now easy to deduce the following properties of $\text{DP}(P)$.

LEMMA 3.1. Suppose $\hat{\omega}_D$ is feasible for $\text{DP}(P)$. If $x(t)$ and $z(t)$ are the piecewise constant extensions of \hat{x} and \hat{z} , respectively, and $y(t)$ is the piecewise linear extension of \hat{y} then $\omega^T(t) = (x^T(t), y^T(t), z^T(t))$ is feasible for SCLP and

$$(3) \quad \int_0^T c^T(t)x(t) dt = \hat{c}_D^T \hat{\omega}_D.$$

Conversely, if $\omega(t)$ is feasible for SCLP with $x(t)$ piecewise constant with partition P , then $x(t_{i-1}+), y(t_i), z(t_{i-1}+), i = 1, \dots, m$, form a feasible solution $\hat{\omega}_D$ for $\text{DP}(P)$ and (3) holds.

3.2. A new discretization. We now introduce the discretization $\text{AP}(P)$.

$$\begin{aligned} \text{AP}(P): \quad & \text{minimize} \quad \sum_{i=1}^m \left(\frac{t_i - t_{i-1}}{2}\right) (c^T(t_{i-1}+)\hat{x}(t_{i-1}+) + c^T(t_i-)\hat{x}(t_i-)) \\ & \text{subject to} \quad \left(\frac{t_1 - t_0}{2}\right) G\hat{x}(t_0+) + \hat{y}\left(\frac{t_1 + t_0}{2}\right) = a\left(\frac{t_1 + t_0}{2}\right), \\ & \quad \left(\frac{t_i - t_{i-1}}{2}\right) G\hat{x}(t_i-) + \hat{y}(t_i) - \hat{y}\left(\frac{t_i + t_{i-1}}{2}\right) \\ & \quad \quad = a(t_i) - a\left(\frac{t_i + t_{i-1}}{2}\right), \quad i = 1, \dots, m, \\ & \quad \left(\frac{t_i - t_{i-1}}{2}\right) G\hat{x}(t_{i-1}+) + \hat{y}\left(\frac{t_i + t_{i-1}}{2}\right) - \hat{y}(t_{i-1}) \\ & \quad \quad = a\left(\frac{t_i + t_{i-1}}{2}\right) - a(t_{i-1}), \quad i = 2, \dots, m, \\ & \quad H\hat{x}(t_{i-1}+) + \hat{z}(t_{i-1}+) = b(t_{i-1}+), \quad i = 1, \dots, m, \\ & \quad H\hat{x}(t_i-) + \hat{z}(t_i-) = b(t_i-), \quad i = 1, \dots, m, \\ & \quad \hat{x}(t_{i-1}+), \hat{x}(t_i-), \hat{y}(t_i), \hat{y}\left(\frac{t_i + t_{i-1}}{2}\right), \\ & \quad \hat{z}(t_{i-1}+), \hat{z}(t_i-) \geq 0, \quad i = 1, \dots, m, \end{aligned}$$

or, in matrix form,

$$\begin{aligned} \text{AP}(P): \quad & \text{minimize} \quad \hat{c}^T \hat{\omega} \\ & \text{subject to} \quad A\hat{\omega} = \hat{b}, \\ & \quad \hat{\omega} \geq 0, \end{aligned}$$

for some appropriate matrix A and vectors \hat{c} and \hat{b} . Again we define for convenience, $\hat{y}(0) = a(0)$. Note that the constraints of $\text{AP}(P)$ are exactly the same as those for $\text{DP}(Q)$, i.e., for the standard discretization with twice as many points (thus the vectors \hat{b} and \hat{c} are twice the dimension of \hat{b}_D and \hat{c}_D). The important difference between AP and DP is in the costs. Instead of taking the average cost on each interval of the partition we take the cost at the end points. We now state an obvious result showing the connections between feasible solutions of $\text{AP}(P)$ and $\text{DP}(P)$.

LEMMA 3.2. *Suppose $\hat{\omega}_D$ is feasible for $\text{DP}(P)$, then by augmenting $\hat{\omega}_D$ using*

$$\begin{aligned} \hat{x}(t_i-) &= \hat{x}(t_{i-1}+), \\ \hat{y}\left(\frac{t_i + t_{i-1}}{2}\right) &= \frac{\hat{y}(t_i) + \hat{y}(t_{i-1})}{2}, \\ \hat{z}(t_i-) &= \hat{z}(t_{i-1}+), \quad \text{for } i = 1, \dots, m, \end{aligned}$$

we obtain a feasible solution $\hat{\omega}$ for $\text{AP}(P)$ with

$$(4) \quad \hat{c}^T \hat{\omega} = \hat{c}_D^T \hat{\omega}_D.$$

Conversely, if $\hat{\omega}$ is feasible for $\text{AP}(P)$ and $\hat{x}(t_{i-1}+) = \hat{x}(t_i-)$ for $i = 1, \dots, m$, then $\hat{x}(t_{i-1}+)$, $\hat{y}(t_i)$, $\hat{z}(t_{i-1}+)$, $i = 1, \dots, m$, form a feasible solution $\hat{\omega}_D$ for $\text{DP}(P)$ and (4) holds.

Now $\text{AP}(P)$ has as its dual

$$\begin{aligned} \text{AP}^*(P): \quad & \text{maximize} \quad \hat{b}^T \hat{\theta} \\ & \text{subject to} \quad \hat{\theta}^T A \leq \hat{c}. \end{aligned}$$

After making the simplifications that $b(t_{i-1}+) = b(t_i-)$ and

$$a\left(\frac{t_i + t_{i-1}}{2}\right) - a(t_{i-1}) = a(t_i) - a\left(\frac{t_i + t_{i-1}}{2}\right),$$

we can write $\text{AP}^*(P)$ as

$$\begin{aligned} \text{AP}^*(P): \quad & \text{maximize} \quad \hat{\pi}^T(t_0+)a(t_0) \\ & \quad + \sum_{i=1}^m (\hat{\pi}(t_{i-1}+) + \hat{\pi}(t_i-))^T \left(a(t_i) - a\left(\frac{t_i + t_{i-1}}{2}\right) \right) \\ & \quad - \sum_{i=1}^m \left(\frac{t_i - t_{i-1}}{2} \right) (\hat{\eta}(t_{i-1}+) + \hat{\eta}(t_i-))^T b(t_i-) \\ \text{subject to} \quad & c(t_i-) - G^T \hat{\pi}(t_i-) + H^T \hat{\eta}(t_i-) \geq 0, \quad i = 1, \dots, m, \\ & c(t_{i-1}+) - G^T \hat{\pi}(t_{i-1}+) + H^T \hat{\eta}(t_{i-1}+) \geq 0, \quad i = 1, \dots, m, \\ & \hat{\eta}(t_i-), \hat{\eta}(t_{i-1}+) \geq 0, \quad i = 1, \dots, m, \\ & \hat{\pi}(t_i-) - \hat{\pi}(t_{i-1}+) \geq 0, \quad i = 1, \dots, m, \\ & \hat{\pi}(t_i+) - \hat{\pi}(t_i-) \geq 0, \quad i = 1, \dots, m-1, \\ & \hat{\pi}(t_m-) \leq 0. \end{aligned}$$

We now present a series of results concerning AP, AP*, SCLP, and SCLP*, most of which can be verified by straightforward algebra, so we will not include all the details in our proofs. The first result shows that we can construct a feasible solution for SCLP* in a natural way from a feasible solution for AP* with the same cost.

THEOREM 3.3. *Suppose $\hat{\theta}$ is feasible for AP*(P), then $\theta^T(t) = (\pi^T(t), \eta^T(t))$, with $\pi(t)$ and $\eta(t)$ the piecewise linear extensions of $\hat{\pi}$ and $\hat{\eta}$, respectively, is feasible for SCLP*. Moreover,*

$$(5) \quad \hat{b}^T \hat{\theta} = - \int_0^T d\pi^T a(t) - \int_0^T \eta^T(t) b(t) dt.$$

Conversely, if $\theta(t)$ is feasible for SCLP with $\pi(t)$ and $\eta(t)$ piecewise linear with partition P , then $\eta(t_{i-1}+)$, $\eta(t_i-)$, $\pi(t_{i-1}+)$, $\pi(t_i-)$, $i = 1, \dots, m$, form a feasible solution $\hat{\theta}$ for AP*(P) and (5) holds.*

Proof. Suppose $\hat{\theta}$ is feasible for AP*(P). By inspection of the inequalities, $\hat{\theta}^T A \leq \hat{c}$ and by the fact that $\pi(t)$ and $\eta(t)$ are piecewise linear with partition P we have $\psi(t)$, $\eta(t) \geq 0$ for all $t \in [0, T]$ and that $\pi(t)$ is monotonic increasing on $[0, T]$ with $\pi(T) = 0$, i.e., that $\theta(t)$ is feasible for SCLP*. Now by the piecewise linearity of $a(t)$ and $\pi(t)$ we have (with the notation that $\pi(t_m+) = \pi(T)$)

$$\begin{aligned} \int_{(t_{i-1}, t_i)} d\pi^T(t) a(t) &= \frac{1}{2} (\pi(t_i-) - \pi(t_{i-1}+))^T (a(t_i) + a(t_{i-1})), \quad i = 1, \dots, m, \\ \int_{\{t_i\}} d\pi^T(t) a(t) &= (\pi(t_i+) - \pi(t_i-))^T a(t_i), \quad i = 1, \dots, m, \\ \int_{\{t_0\}} d\pi^T(t) a(t) &= 0. \end{aligned}$$

Also as $\eta(t)$ is piecewise linear and $b(t)$ is piecewise constant with partition P we have

$$\int_{t_{i-1}}^{t_i} \eta^T(t) b(t) dt = \left(\frac{t_i - t_{i-1}}{2} \right) (\hat{\eta}(t_{i-1}+) + \hat{\eta}(t_i-))^T b(t_i-), \quad i = 1, \dots, m.$$

The proof that the objective functions of the two linear programs is the same is now a straightforward algebraic exercise.

The proof of the converse is similar. \square

The next result is the analogous result for AP(P) and SCLP. However, the result is not as strong in that a feasible solution for SCLP gives one for AP(P) of the same cost but not all feasible solutions for AP(P) correspond to feasible solutions for SCLP.

THEOREM 3.4. *Suppose $\omega(t)$ is feasible for SCLP with $x(t)$ piecewise constant with partition P . Then $x(t_{i-1}+)$, $x(t_i-)$, $y(t_i)$, $z(t_{i-1}+)$, $z(t_i-)$, $i = 1, \dots, m$, form a feasible solution $\hat{\omega}$ for AP(P). Moreover,*

$$(6) \quad \hat{c}^T \hat{\omega} = \int_0^T c^T(t) x(t) dt.$$

Conversely, if $\hat{\omega}$ is feasible for AP(P) and $\hat{x}(t_{i-1}+) = \hat{x}(t_i-)$ for $i = 1, \dots, m$, then the $\omega(t)$ formed by taking $y(t)$ as the piecewise linear extension of \hat{y} and $x(t)$ and $z(t)$ as the piecewise constant extensions of \hat{x} and \hat{z} , respectively, is feasible for SCLP and (6) holds.

Proof. The result follows immediately from Lemmas 3.1 and 3.2. \square

Using Theorem 3.3 we can now show, by using the strong duality theorem for finite-dimensional linear programming (see, e.g., Dantzig [9]), that the optimal value of the problem $AP(P)$ constructed from a feasible solution for SCLP with $x(t)$ piecewise constant with partition P , provides a lower bound on the optimal value of SCLP.

THEOREM 3.5. *Suppose $\omega(t)$ is feasible for SCLP with $x(t)$ piecewise constant with partition P . Then*

$$V[AP(P)] \leq V[SCLP^*] \leq V[SCLP].$$

Proof. See Fig. 1. By Theorem 3.3 and the duality theorem for finite-dimensional linear programming we have $V[AP(P)] = V[AP^*(P)] \leq V[SCLP^*]$. The result now follows by Lemma 2.1. \square

COROLLARY 3.6. *Suppose $\omega(t)$ is feasible for SCLP and that the corresponding $\hat{\omega}$ is optimal for $AP(P)$, then $\omega(t)$ is optimal for SCLP.*

Proof. By Theorem 3.4 we have

$$V[SCLP] \leq \int_0^T c^T(t)x(t) dt = V[AP(P)],$$

and so the result follows by Theorem 3.5. \square

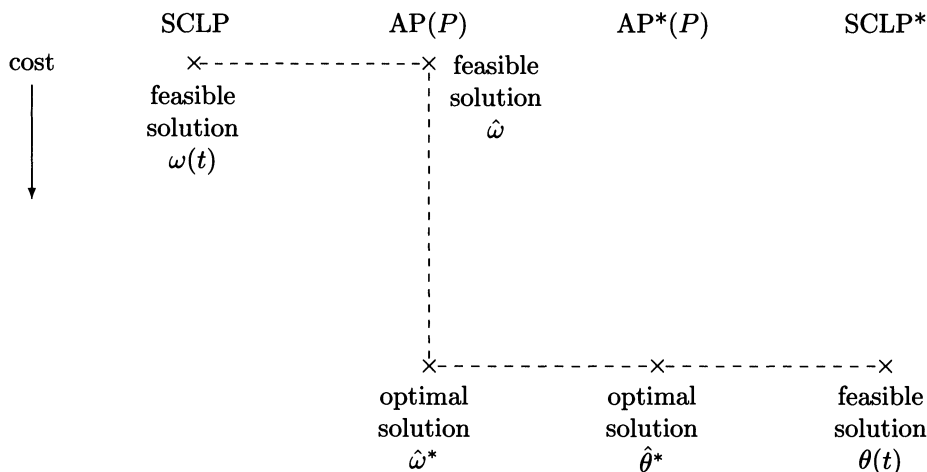


FIG. 1. Cost relationship between solutions of SCLP, $AP(P)$, $AP^*(P)$ and $SCLP^*$.

The next step is to construct a feasible solution for SCLP which in some way is similar to a given one for $AP(P)$. To do this we extend the solution for $AP(P)$ across each interval of the partition making it piecewise constant with partition Q . Suppose $\hat{\omega}$ is feasible for $AP(P)$. Define

$$x(t) = \begin{cases} \hat{x}(t_{i-1}+), & t \in \left[t_{i-1}, \frac{t_i + t_{i-1}}{2} \right), \quad i = 1, \dots, m, \\ \hat{x}(t_i-), & t \in \left[\frac{t_i + t_{i-1}}{2}, t_i \right), \quad i = 1, \dots, m, \\ \hat{x}(t_m-), & t = T. \end{cases}$$

We then define $y(t)$ and $z(t)$ from the constraints of SCLP (i.e., satisfying (1) and (2)) and set

$$\omega^T(t) = (x^T(t), y^T(t), z^T(t)).$$

Finally we let

$$(7) \quad \alpha(\omega) = \hat{c}^T \hat{\omega} - \int_0^T c^T(t)x(t) dt.$$

The next theorem gives the properties of $\omega(t)$.

THEOREM 3.7. $\omega(t)$ is feasible for SCLP. Moreover,

$$\alpha(\omega) = \sum_{i=1}^m \left(\frac{(t_i - t_{i-1})^2}{8} \right) (x(t_i-) - x(t_{i-1}+))^T \dot{c}(t_i-).$$

Proof. It is not difficult to show that $y(t)$ is the piecewise linear extension of \hat{y} in $\hat{\omega}$ and that

$$z(t) = \begin{cases} \hat{z}(t_{i-1}+), & t \in \left[t_{i-1}, \frac{t_i + t_{i-1}}{2} \right), \quad i = 1, \dots, m, \\ \hat{z}(t_i-), & t \in \left[\frac{t_i + t_{i-1}}{2}, t_i \right), \quad i = 1, \dots, m, \\ \hat{z}(t_m), & t = T, \end{cases}$$

and so from this it is clear that $\omega(t)$ is feasible for SCLP. The cost relationship follows after simple integration and algebra. \square

Finally we consider what happens when the partition is made finer and finer. From Theorem 3.7 we are now able to deduce that, under Assumption 2, as the norm of the partition tends to zero, the values of the linear programs SCLP and AP tend to the same value. This gives the result that $V[\text{SCLP}] = V[\text{SCLP}^*]$ as an immediate consequence, however this result will be strengthened in the next section to show that SCLP* actually attains $V[\text{SCLP}^*]$ (Corollary 4.5). We first give an additional lemma about sequences.

LEMMA 3.8. Suppose $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ are sequences such that

$$b_n - a_m \geq 0, \quad m, n = 1, 2, \dots, \\ \lim_{n \rightarrow \infty} (b_n - a_n) = 0.$$

then $\lim_{n \rightarrow \infty} a_n$ and $\lim_{n \rightarrow \infty} b_n$ exist and

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n.$$

Proof. Let

$$\varepsilon = \inf_n b_n - \sup_n a_n,$$

then $\varepsilon \geq 0$. Now since $b_n - a_n \geq \varepsilon$ for every n we must have $\varepsilon = 0$. Let $L = \inf_n b_n = \sup_n a_n$. Now since $a_m \leq L \leq b_n$ for all m ,

$$(L - a_n) \leq (b_n - a_n) \quad \text{and} \quad (b_n - L) \leq (b_n - a_n),$$

and so the result follows. \square

COROLLARY 3.9. Suppose Assumption 2 holds and that $\omega(t)$ is feasible for SCLP with $x(t)$ piecewise constant with partition P , then

$$(8) \quad \lim_{\substack{\|Q\| \rightarrow 0 \\ P \subseteq Q}} V[\text{AP}(Q)] = \lim_{\substack{\|Q\| \rightarrow 0 \\ P \subseteq Q}} \int_0^T c^T(t)x_Q(t) dt,$$

where $x_Q(t)$ is obtained from the optimal solution for $\text{AP}(Q)$ as outlined above, and hence

$$(9) \quad \lim_{\substack{\|Q\| \rightarrow 0 \\ P \subseteq Q}} \int_0^T c^T(t) x_Q(t) dt = V[\text{SCLP}] = V[\text{SCLP}^*].$$

Proof. Suppose for any feasible $x(t)$, $\|x(t)\| \leq M$. Also suppose that $\|\dot{c}(t)\| \leq C$ for $t \in [0, T] - P$. Let $Q = \{s_1, \dots, s_k\} \supseteq P$. Then

$$|\alpha(\omega_Q)| \leq \frac{MC}{4} \sum_{i=1}^k (s_i - s_{i-1})^2 \leq \frac{MC\|Q\|T}{4},$$

and so $\alpha(\omega_Q) \rightarrow 0$ as $\|Q\| \rightarrow 0$.

Let $\{Q_n\}_{n=1}^\infty$ be any sequence of partitions with $\lim_{n \rightarrow \infty} \|Q_n\| = 0$ and $P \subseteq Q_n$. Let

$$a_n = V[\text{AP}(Q_n)], \quad \text{and} \quad b_n = \int_0^T c^T(t) x_{Q_n}(t) dt,$$

then by the above we see that $\lim_{n \rightarrow \infty} (b_n - a_n) = 0$. Also by Theorem 3.5 we have $b_n - a_m \geq 0$ for all m and n and so (8) now follows by Lemma 3.8. We now have (9) again by Theorem 3.5. \square

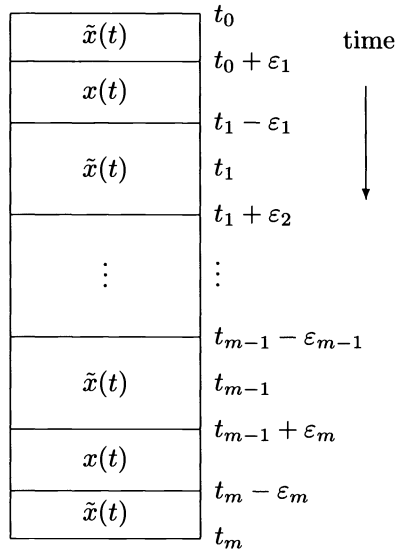
4. Constructing an improved solution. In this section we show how to construct an improved solution for SCLP, if one exists, from some starting feasible solution for SCLP. This will form the basis of an algorithm for solving SCLP. To do this we use the results the previous section. We will again fix the partition P as in the previous section and any reference to AP will denote $\text{AP}(P)$.

We begin by giving an outline of our approach. Any feasible solution for SCLP with $x(t)$ piecewise constant with partition P generates a feasible solution for $\text{AP}(P)$ (Theorem 3.4). If the finite-dimensional program AP is not optimal it is possible to find a solution for AP with strictly improved objective function. From the previous section we are able to construct a feasible solution for SCLP based on this solution in a fairly natural way. We then “patch” this solution together with our old one to produce another solution whose x value is the same as the original one in some time intervals and the same as the one constructed from AP otherwise. If this is done in an appropriate way, we obtain a new feasible solution for SCLP with strictly improved objective function. If, however, the current SCLP solution is optimal for AP, then by Corollary 3.6, our current solution is optimal for SCLP.

Let $\omega^T(t) = (x^T(t), y^T(t), z^T(t))$ be feasible for SCLP with $x(t)$ piecewise constant with partition P . Using the notation of the previous section, let $\hat{\omega}$ denote the corresponding solution for AP constructed from the partition P . Suppose $\hat{\omega}$ is not optimal for AP. Then $\hat{\hat{\omega}}$ exists, feasible for AP, with strictly improved objective function, i.e.,

$$\delta \stackrel{\text{def}}{=} c^T \hat{\hat{\omega}} - c^T \hat{\omega} < 0.$$

Let $\tilde{\omega}^T(t) = (\tilde{x}^T(t), \tilde{y}^T(t), \tilde{z}^T(t))$ be the feasible solution for SCLP generated by $\hat{\hat{\omega}}$ as defined in the previous section. Then $\tilde{\omega}(t)$ is constant on $[t_{i-1}, (t_{i-1} + t_i)/2)$ and on $[(t_{i-1} + t_i)/2, t_i]$ for $i = 1, \dots, m$.

FIG. 2. Construction of $\bar{x}_\varepsilon(t)$.

We now “patch” together $\omega(t)$ and $\tilde{\omega}(t)$ as follows (see Fig. 2). Let

$$\tau = \min_{i=1, \dots, m} \left(\frac{t_i - t_{i-1}}{2} \right),$$

and $\varepsilon \in [0, \tau]$. Define

$$\varepsilon_i = \frac{(t_i - t_{i-1})\varepsilon}{2\tau}.$$

Note that if $\varepsilon = \tau$ then $\varepsilon_i = (t_i - t_{i-1})/2$. Define

$$\bar{x}_\varepsilon(t) = \begin{cases} \tilde{x}(t), & t \in [t_{i-1}, t_{i-1} + \varepsilon_i) \cup [t_i - \varepsilon_i, t_i), \quad i = 1, \dots, m, \\ x(t), & \text{otherwise.} \end{cases}$$

Let $\bar{y}_\varepsilon(t)$ and $\bar{z}_\varepsilon(t)$ be given by the constraints of SCLP, i.e.,

$$\begin{aligned} \int_0^t G\bar{x}_\varepsilon(s) ds + \bar{y}_\varepsilon(t) &= a(t), \\ H\bar{x}_\varepsilon(t) + \bar{z}_\varepsilon(t) &= b(t), \quad t \in [0, T]. \end{aligned}$$

We now claim that $\bar{\omega}_\varepsilon^T(t) = (\bar{x}_\varepsilon^T(t), \bar{y}_\varepsilon^T(t), \bar{z}_\varepsilon^T(t))$ is feasible for SCLP.

Clearly $\bar{x}_\varepsilon(t) \geq 0$ for all $t \in [0, T]$ as $x(t), \tilde{x}(t) \geq 0$ for all $t \in [0, T]$ by the feasibility of $\omega(t)$ and $\tilde{\omega}(t)$. Also for $t \in [t_{i-1} + \varepsilon_i, t_i - \varepsilon_i)$, $\bar{x}_\varepsilon(t) = x(t)$, so $\bar{z}_\varepsilon(t) = z(t)$ for $t \in [t_{i-1} + \varepsilon_i, t_i - \varepsilon_i)$ and for other t , $\bar{x}_\varepsilon(t) = \tilde{x}(t)$, and so $\bar{z}_\varepsilon(t) = \tilde{z}(t)$. Hence $\bar{z}_\varepsilon(t) \geq 0$ for all $t \in [0, T]$, again by the feasibility of $\omega(t)$ and $\tilde{\omega}(t)$.

Thus to establish the feasibility of $\bar{\omega}_\varepsilon(t)$ it remains to show that $\bar{y}_\varepsilon(t) \geq 0$. As $\bar{x}_\varepsilon(t)$ is constant on the intervals $[t_{i-1}, t_{i-1} + \varepsilon_i)$, $[t_{i-1} + \varepsilon_i, t_i - \varepsilon_i)$ and $[t_i - \varepsilon_i, t_i)$ for $i = 1, \dots, m$, $\bar{y}_\varepsilon(t)$ is linear on these intervals, so it is sufficient to show that $\bar{y}_\varepsilon(t_{i-1})$, $\bar{y}_\varepsilon(t_{i-1} + \varepsilon_i)$, $\bar{y}_\varepsilon(t_i - \varepsilon_i)$ and $\bar{y}_\varepsilon(t_i) \geq 0$ for $i = 1, \dots, m$. This is the content of the next lemma.

LEMMA 4.1.

$$(10) \quad \bar{y}_\varepsilon(t_{i-1} + \varepsilon_i) = \left(1 - \frac{\varepsilon}{\tau}\right) y(t_{i-1}) + \frac{\varepsilon}{\tau} \tilde{y}\left(\frac{t_i + t_{i-1}}{2}\right),$$

$i = 1, \dots, m,$

$$(11) \quad \bar{y}_\varepsilon(t_i - \varepsilon_i) = \left(1 - \frac{\varepsilon}{\tau}\right) y(t_i) + \frac{\varepsilon}{\tau} \tilde{y}\left(\frac{t_i + t_{i-1}}{2}\right),$$

$i = 1, \dots, m,$

$$(12) \quad \bar{y}_\varepsilon(t_i) = \left(1 - \frac{\varepsilon}{\tau}\right) y(t_i) + \frac{\varepsilon}{\tau} \tilde{y}(t_i),$$

$i = 0, \dots, m.$

Proof. We establish these equalities by induction. Clearly $\bar{y}_\varepsilon(0) = \tilde{y}(0) = y(0) = a(0)$ so (12) is true for $i = 0$. Assume (12) is true for $i - 1$. We show that (10)–(12) are true for i .

Now $\bar{x}_\varepsilon(t) = \tilde{x}(t)$ on $[t_{i-1}, t_{i-1} + \varepsilon_i]$, so

$$\begin{aligned} \bar{y}_\varepsilon(t_{i-1} + \varepsilon_i) - \bar{y}_\varepsilon(t_{i-1}) &= \tilde{y}(t_{i-1} + \varepsilon_i) - \tilde{y}(t_{i-1}) \\ &= \frac{2\varepsilon_i}{t_i - t_{i-1}} \left[\tilde{y}\left(\frac{t_i + t_{i-1}}{2}\right) - \tilde{y}(t_{i-1}) \right], \end{aligned}$$

as \tilde{y} is linear on $\left[t_{i-1}, \frac{t_i + t_{i-1}}{2}\right]$. Hence

$$\bar{y}_\varepsilon(t_{i-1} + \varepsilon_i) - \bar{y}_\varepsilon(t_{i-1}) = \frac{\varepsilon}{\tau} \left[\tilde{y}\left(\frac{t_i + t_{i-1}}{2}\right) - \tilde{y}(t_{i-1}) \right].$$

So by the assumption that (12) holds for $i - 1$,

$$\bar{y}_\varepsilon(t_{i-1} + \varepsilon_i) = \left(1 - \frac{\varepsilon}{\tau}\right) y(t_{i-1}) + \frac{\varepsilon}{\tau} \tilde{y}\left(\frac{t_i + t_{i-1}}{2}\right),$$

i.e., (10) is satisfied for i .

Similarly, $\bar{x}_\varepsilon(t) = x(t)$ on $[t_{i-1} + \varepsilon_i, t_i - \varepsilon_i]$, so

$$\begin{aligned} \bar{y}_\varepsilon(t_i - \varepsilon_i) - \bar{y}_\varepsilon(t_{i-1} + \varepsilon_i) &= y(t_i - \varepsilon_i) - y(t_{i-1} + \varepsilon_i) \\ &= \left(1 - \frac{2\varepsilon_i}{t_i - t_{i-1}}\right) (y(t_i) - y(t_{i-1})), \end{aligned}$$

again as y is linear on $[t_{i-1}, t_i]$. Hence as (10) is true for i so is (11).

A similar argument shows that (12) is true for i and hence the lemma is established by induction. \square

COROLLARY 4.2. $\bar{\omega}_\varepsilon$ is feasible for all $\varepsilon \in [0, \tau]$.

Proof. We just need to establish that $\bar{y}_\varepsilon(t) \geq 0$ for all $t \in [0, T]$. Now $\bar{y}_\varepsilon(t_i)$ is a convex combination of $y(t_i)$ and $\tilde{y}(t_i)$ and so as $\omega(t)$ and $\tilde{\omega}(t)$ are feasible for SCLP, $y(t_i), \tilde{y}(t_i) \geq 0$. Hence $\bar{y}_\varepsilon(t_i) \geq 0$ for $i = 0, \dots, m$. Similarly, $\bar{y}_\varepsilon(t_{i-1} + \varepsilon_i), \bar{y}_\varepsilon(t_i - \varepsilon_i) \geq 0$ for $i = 1, \dots, m$. The result now follows from the piecewise linearity of \bar{y}_ε . \square

Having established the feasibility of $\bar{\omega}_\varepsilon$ the next step is to show that for small enough ε an improvement in the objective function can always be made. This is the content of the next lemma and following corollary.

LEMMA 4.3.

$$\int_0^T c^T(t) \bar{x}_\varepsilon(t) dt - \int_0^T c^T(t) x(t) dt = \frac{\varepsilon}{\tau} \left(\delta - \frac{\varepsilon \alpha}{\tau} \right),$$

where $\alpha = \alpha(\tilde{\omega})$ as given by (7).

Proof. As $c(t)$ is linear on (t_{i-1}, t_i) for $i = 1, \dots, m$, we have

$$\begin{aligned} \int_{t_{i-1}}^{t_{i-1}+\varepsilon_i} c(t) dt &= \frac{\varepsilon_i}{2} (c(t_{i-1}+) + c(t_{i-1} + \varepsilon_i)) \\ &= \varepsilon_i c(t_{i-1}+) + \frac{\varepsilon_i^2 (t_i - t_{i-1})^2}{8\tau^2} \dot{c}(t_{i-1}). \end{aligned}$$

Similarly,

$$\int_{t_i-\varepsilon_i}^{t_i} c(t) dt = \varepsilon_i c(t_i-) - \frac{\varepsilon_i^2 (t_i - t_{i-1})^2}{8\tau^2} \dot{c}(t_i-).$$

Now for $t \in [0, T)$,

$$\bar{x}_\varepsilon(t) - x(t) = \begin{cases} \tilde{x}(t_{i-1}+) - x(t_{i-1}+), & t \in [t_{i-1}, t_{i-1} + \varepsilon_i), \quad i = 1, \dots, m, \\ \tilde{x}(t_i-) - x(t_i-), & t \in [t_i - \varepsilon_i, t_i), \quad i = 1, \dots, m, \\ 0, & \text{otherwise.} \end{cases}$$

So

$$\begin{aligned} &\int_0^T c^T(t) \bar{x}_\varepsilon(t) dt - \int_0^T c^T(t) x(t) dt \\ &= \sum_{i=1}^m \int_{t_{i-1}}^{t_{i-1}+\varepsilon_i} c^T(t) (\tilde{x}(t) - x(t)) dt + \sum_{i=1}^m \int_{t_i-\varepsilon_i}^{t_i} c^T(t) (\tilde{x}(t) - x(t)) dt \\ &= \sum_{i=1}^m (\tilde{x}(t_{i-1}+) - x(t_{i-1}+))^T \left(\varepsilon_i c(t_{i-1}+) + \frac{\varepsilon_i^2 (t_i - t_{i-1})^2}{8\tau^2} \dot{c}(t_{i-1}) \right) \\ &\quad + \sum_{i=1}^m (\tilde{x}(t_i-) - x(t_i-))^T \left(\varepsilon_i c(t_i-) - \frac{\varepsilon_i^2 (t_i - t_{i-1})^2}{8\tau^2} \dot{c}(t_i-) \right). \end{aligned}$$

The result now follows after a little algebra and the fact that

$$\begin{aligned} \delta &= \hat{c}^T \hat{\tilde{\omega}} - \hat{c}^T \hat{\omega} \\ &= \sum_{i=1}^m \left(\frac{t_i - t_{i-1}}{2} \right) (\tilde{x}(t_{i-1}+) - x(t_{i-1}+))^T c(t_{i-1}+) \\ &\quad + \sum_{i=1}^m \left(\frac{t_i - t_{i-1}}{2} \right) (\tilde{x}(t_i-) - x(t_i-))^T c(t_i-). \quad \square \end{aligned}$$

COROLLARY 4.4. For ε sufficiently small, $\int_0^T c^T(t) \bar{x}_\varepsilon(t) dt < \int_0^T c^T(t) x(t) dt$ and

$$\min_{\varepsilon} \int_0^T c^T(t) \bar{x}_\varepsilon(t) dt - \int_0^T c^T(t) x(t) dt = \begin{cases} \frac{\delta^2}{4\alpha}, & \alpha < 0 \text{ and } \frac{\delta}{2\alpha} < 1, \\ \delta - \alpha, & \text{otherwise,} \end{cases}$$

and occurs at

$$(13) \quad \varepsilon^* = \begin{cases} \frac{\delta\tau}{2\alpha}, & \alpha < 0 \text{ and } \frac{\delta}{2\alpha} < 1, \\ \tau, & \text{otherwise.} \end{cases}$$

We now have a strong duality result. This is the main result in Tyndall [19] proved for the dual SCLP* and via a sequence of discretizations of the form $DP(P_n)$. However, while Tyndall's result is not restricted to problems with piecewise constant $x(t)$, it is restricted to positive G , H , $a(t)$, and $b(t)$. The result was later strengthened by a number of authors, e.g., Grinold [12] and Tyndall [20] but again there is some restriction on the form of the problem.

COROLLARY 4.5 (Strong duality). *Suppose Assumption 2 holds. If $\omega(t)$ is optimal for SCLP then there exists $\theta(t)$ optimal for SCLP* with*

$$\int_0^T c^T(t)x(t) dt = - \int_0^T d\pi^T(t)a(t) - \int_0^T \eta^T(t)b(t) dt.$$

Proof. If $\omega(t)$ is optimal for SCLP then we can assume that $x(t)$ is piecewise constant by a previous remark. The corresponding $\hat{\omega}$ is then optimal for AP, otherwise we can construct an improved solution for SCLP as outlined above. The result now follows by Corollary 3.6. \square

5. Implementations of the algorithm and computational results. We now have all the ingredients for implementing an algorithm to solve SCLP. In this section we outline different possible ways of putting these ingredients together and discuss computational results from one such implementation.

1. *Pure discretization.* Solve $AP(P)$ for some partition P and construct a feasible solution for SCLP as outlined in §3.2. This has two advantages over just solving $DP(P)$. First, by Theorem 3.5, the cost of the AP solution provides a lower bound on the cost of the optimal solution for SCLP. Second, if we have Assumption 2, then by using Theorem 3.7 and Corollary 3.9 we can construct a partition so that the SCLP solution will be within any desired amount of the optimal solution.

2. *Continuous time.* Given a feasible solution for SCLP, construct an improvement as outlined in §4, if possible, repeating until the solution is within a prescribed limit. This has several variations, some of which we outline below.

- By Lemma 3.1, a solution for SCLP generates one for $DP(P)$ for some P at the same cost. $DP(P)$ can then be solved, or at least improved, and this solution used as the starting feasible SCLP solution from which an improved SCLP solution is obtained along the lines of §4. Solving $DP(P)$ completely will produce the best SCLP solution for the partition P .

- By Theorem 3.4, a solution for SCLP generates one for $AP(P)$ for some P at the same cost. Solving $AP(P)$ to completion and using this as the improved $AP(P)$ solution for patching together with the old one also produces a lower bound on the objective function, as was shown in the discrete algorithm above.

- Given the SCLP solution, we construct the AP solution, purify this (i.e., construct a basic feasible solution) and perform one (or only a few) simplex pivots. This is the improved AP solution, which, after patching together with the old one, produces an improved SCLP solution.

It is still not clear which of these implementations will produce the best results. However, one simple change that can be applied to each implementation above does

seem to improve the solution time considerably, as well as improving the convergence of the partition points. This change involves removing unwanted partition points. After completing each iteration of the algorithm we check to see if the new SCLP solution (as well as $\dot{a}(t)$, $b(t)$, and $\dot{c}(t)$) is constant across two or more intervals of the new partition. If it is, then we remove partition points so that there is only one such interval. This will produce a new partition with fewer points and with each point corresponding to a discontinuity in one or more components of $x(t)$, $\dot{a}(t)$, $b(t)$ or $\dot{c}(t)$.

Before proceeding to give computational results we give a brief discussion of the algorithm for networks given in Anderson and Philpott [4]. In this paper the authors give an optimality test followed by a discussion of a simplex-like algorithm for the problem by showing how to perform a pivot when one of the conditions of the optimality test is violated. Two different types of pivot are performed depending on which optimality conditions are violated. These two conditions are called first- and second-order suboptimality, with most pivots arising from the first-order suboptimality case. It can be shown that their optimality test is no more than a statement of our weak duality theorem (Lemma 2.1). Moreover, the pivot for the first-order suboptimality case also corresponds to one possible implementation of our algorithm. We describe this pivot in the context of our SCLP algorithm. The pivot starts off with a basic feasible solution for SCLP (in this case CNP), which is also basic for $AP(P)$ for some P . One simplex pivot is then performed on $AP(P)$ and the solution patched together with the old one to produce a new feasible solution for CNP. A complicated purification step then follows to produce a new basic feasible solution for CNP.

The pivot arising from the second-order suboptimality case, while not one of our implementations, is very similar and so further improvements to the algorithm for CNP could be made by considering the SCLP algorithm discussed here.

The method we use in the examples below is no doubt inefficient but has the advantage of being easy to implement. It follows along the lines of a combination of the first two continuous-time examples given above. Given a partition P , we solve $DP(P)$ to optimality. $AP(P)$ is then solved to optimality and the objective function then provides a bound on the optimal objective function for SCLP. We then combined the DP and AP solutions using the method described in §4 with ε given by (13), to produce a new partition. Finally, unwanted partition points are removed to produce the partition Q . This was the new partition for the next step.

We now present results from solving two simple problems. The first is a network problem and has been solved in Anderson and Philpott [5] using their algorithm for networks. The second is taken from Anderson [1].

Problem 1. $T = 10$, $n_1 = 5$, $n_2 = 8$, $n_3 = 5$.

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

$$a_1(t) = \begin{cases} 4t, & t \in [0, 5], \\ 2(t+5), & t \in [5, 10]. \end{cases} \quad a_2(t) = \begin{cases} t, & t \in [0, 5], \\ 2t-5, & t \in [5, 10]. \end{cases}$$

$$a_3(t) = \begin{cases} -t, & t \in [0, 5], \\ 5-2t, & t \in [5, 10]. \end{cases} \quad a_4(t) = -3t, \\ a_i(t) = \infty, \text{ for } i = 5, 6, 7, 8.$$

$$b_i(t) = \begin{cases} 1, & i = 3, \\ 2, & \text{otherwise.} \end{cases}$$

$$c_1(t) = 10 - \frac{3}{5}t, \quad c_2(t) = 7, \\ c_3(t) = 6 - \frac{3}{5}t, \quad c_4(t) = 4, \\ c_5(t) = 2 + t.$$

The optimal solution is

$$x_1^*(t) = \begin{cases} 1, & t \in [0, 3.75), \\ 2, & t \in [3.75, 5), \\ 1, & t \in [5, 8.75), \\ 0, & t \in [8.75, 10]. \end{cases} \quad x_2^*(t) = 2.$$

$$x_3^*(t) = \begin{cases} 0, & t \in [0, 3.75), \\ 1, & t \in [3.75, 10]. \end{cases} \quad x_4^*(t) = \begin{cases} 2, & t \in [0, 8.75), \\ 1, & t \in [8.75, 10]. \end{cases}$$

$$x_5^*(t) = \begin{cases} 1, & t \in [0, 8.75), \\ 2, & t \in [8.75, 10]. \end{cases}$$

and has objective function 396.25.

An initial partition consisting of the points where the functions $a(t)$, $b(t)$ and $c(t)$ change was chosen, i.e., $P_0 = \{0, 5, 10\}$. The results are given in Table 1. At the beginning of each iteration we start off with the partition given in the partition column. DP is then solved for this partition and the cost of this (feasible SCLP) solution is given. Then AP is solved and the cost of this (feasible SCLP*) solution is given. The cost of the SCLP solution constructed from this AP solution is then given, followed by the cost of the solution obtained from patching together this solution with the DP solution. It is worth noting that the cost obtained its optimal value at the

TABLE 1

Iteration	Partition	DP	AP	SCLP	Combined
1	{0,5,10}	397.5	392.5	397.5	396.25
2	{0,3.75,5,8.75,10}	396.25	396.25	396.5625	-

end of the first iteration and in fact the combined solution was exactly $x^*(t)$ given above. However, it was not until an SCLP* solution with the same cost was obtained that the algorithm was able to detect optimality. Note also that the SCLP solution constructed from the AP one on the second iteration was not the optimal one.

Problem 2. $T = 2$, $n_1 = 2$, $n_2 = 2$, $n_3 = 1$.

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 2 \end{bmatrix}.$$

$a_1(t) = 4 + t, \quad a_2 = 3 + 2t.$

$b_1(t) = 10.$

$c_1(t) = t - 2, \quad c_2(t) = t - 2.$

The optimal solution is

$$x_1^*(t) = \begin{cases} 10, & t \in [0, \frac{4}{9}), \\ 1, & t \in [\frac{4}{9}, 2]. \end{cases} \quad x_2^*(t) = \begin{cases} 0, & t \in [0, \frac{4}{9}), \\ \frac{9}{2}, & t \in [\frac{4}{9}, 2]. \end{cases}$$

and has objective function $-14\frac{5}{9}$.

This example has the advantage that the algorithm can be seen to converge to the optimum rather than obtain it explicitly as is the case of the first example.

Again an initial partition consisting of the points where the functions $a(t)$, $b(t)$, and $c(t)$ change was chosen, i.e., $P_0 = \{0, 2\}$. The results are presented in Tables 2 and 3.

TABLE 2

Iteration	Partition
1	{0,2}
2	{0,1,2}
3	{0,0.5,2}
4	{0,0.25,0.5,2}
5	{0,0.375,0.5,2}
6	{0,0.4375,0.5,2}
7	{0,0.4375,0.46875,2}
8	{0,0.4375,0.453125,2}
9	{0,0.4375,0.4453124985,2}

TABLE 3

Iteration	DP	AP	SCLP	Combined
1	-13	-15	-14	-14
2	-14	-15	-14.5	-14.5
3	-14.5	-14.5625	-14.53125	-14.53125
4	-14.53125	-14.5625	-14.546875	-14.546875
5	-14.546875	-14.5625	-14.5546875	-14.5546875
6	-14.5546875	-14.555664063	-14.555175781	-14.555175781
7	-14.555175781	-14.555664063	-14.555419922	-14.555419922
8	-14.555419922	-14.555664063	-14.555541992	-14.555541992
9	-14.555541992	-14.555557251	-14.555549622	-

The algorithm stopped at the ninth iteration as the difference in cost between the SCLP solution and that for SCLP* was within the prescribed limit of 10^{-5} . The final SCLP solution was

$$x_1(t) = \begin{cases} 10, & t \in [0, 0.4414062492), \\ 8.0000038147, & t \in [0.4414062492, 0.4453124985), \\ 1, & t \in [0.4453124985, 2]. \end{cases}$$
$$x_2(t) = \begin{cases} 0, & t \in [0, 0.4414062492), \\ 0.9999982715, & t \in [0.4414062492, 0.4453124985), \\ 4.5, & t \in [0.4453124985, 2]. \end{cases}$$

Several points are worth mentioning. First, it can be seen from the above example that in deleting partition points where the SCLP solution is constant, care needs to be taken to avoid numerical problems. This is because the intervals of the partition are then of very different sizes, thus giving rise to very different sized numbers in the \hat{b} or \hat{b}_D . The problem should be easily overcome by scaling the constraints, or better, by replacing a constraint such as

$$(t_i - t_{i-1})G\hat{x}(t_{i-1}+) + \hat{y}(t_i) - \hat{y}(t_{i-1}) = a(t_i) - a(t_{i-1}),$$

in $DP(P)$ by

$$G\hat{x}(t_{i-1}+) + \hat{y}(t_{i-1}+) = \dot{a}(t_{i-1}+),$$

with similar changes in $AP(P)$.

Second, it has been observed that the results obtained in this example are very sensitive to the way AP or DP are solved. For example, by changing the way an initial basic feasible solution was chosen for either AP or DP , it was possible to produce results where there was a very large partition size with many points clustered around $4/9$. Another change lead to a partition size of only three, with the middle value approaching $4/9$. However, in all cases, convergence of the cost was observed. The reason for this is that because $c_1(t)$ and $c_2(t)$ are the same, it is possible to produce different solutions with the same cost, which could lead to solutions with a differing number of partition points. This effect could then be compounded over several iterations. This is a type of degeneracy and as with degeneracy in the finite-dimensional setting, this problem can be overcome by slightly perturbing the two costs.

Acknowledgments. The author thanks Dr. E. J. Anderson for his many helpful suggestions and discussions and Trinity College, Cambridge for providing funds to make this research possible.

REFERENCES

- [1] E. J. ANDERSON, *A Continuous Model For Job-Shop Scheduling*, Ph. D. thesis, University of Cambridge, Cambridge, U.K., 1978.
- [2] E. J. ANDERSON, P. NASH, AND A. F. PEROLD, *Some properties of a class of continuous linear programs*, SIAM J. Control Optim., 21 (1983), pp. 758–765.
- [3] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite Dimensional Spaces*, Wiley-Interscience, Chichester, 1987.
- [4] E. J. ANDERSON AND A. B. PHILPOTT, *A continuous-time network simplex algorithm*, Networks, 19 (1989), pp. 395–425.
- [5] ———, *Erratum: a continuous-time network simplex algorithm*, Networks, 19 (1989), pp. 823–827.
- [6] K. M. ANSTREICHER, *Generation of feasible descent directions in continuous-time linear programming*, Tech. Report SOL 83-18, Department of Operations Research, Stanford University, Stanford, CA, 1983.
- [7] R. E. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
- [8] R. N. BUIE AND J. ABRHAM, *Numerical solutions to continuous linear programming problems*, Z. Oper. Res., 17 (1973), pp. 107–117.
- [9] G. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
- [10] W. P. DREWS, *A simplex-like algorithm for continuous-time linear optimal control problems*, in Optimization Methods for Resource Allocation, R.W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 309–322.
- [11] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part I: General Theory*, Wiley-Interscience, New York, 1988.

- [12] R. C. GRINOLD, *Continuous programming part one: linear objectives*, J. Math. Anal. Appl., 28 (1969), pp. 32–51.
- [13] R. J. HARTBERGER, *Representation extended to continuous time*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 297–307.
- [14] R. S. LEHMAN, *On the continuous simplex method*, RM-1386, Rand Corporation, Santa Monica, CA, 1954.
- [15] N. LEVINSON, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73–83.
- [16] N. S. PAPAGEORGIOU, *A class of infinite dimensional linear programming problems*, J. Math. Anal. Appl., 87 (1982), pp. 228–245.
- [17] A. F. PEROLD, *Fundamentals of a continuous time simplex method*, Tech. Report SOL 78-26, Department of Operations Research, Stanford University, Stanford, CA, 1978.
- [18] R. G. SEGERS, *A generalised function setting for dynamic optimal control problems*, in Optimization Methods for Resource Allocation, R.W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 279–296.
- [19] W. F. TYNDALL, *A duality theorem for a class of continuous linear programming problems*, SIAM J. Appl. Math., 13 (1965), pp. 644–666.
- [20] ———, *An extended duality theorem for continuous linear programming problems*, SIAM J. Appl. Math., 15 (1967), pp. 1294–1298.

A PATH-FOLLOWING ALGORITHM FOR LINEAR PROGRAMMING USING QUADRATIC AND LOGARITHMIC PENALTY FUNCTIONS*

PAUL TSENG†

Abstract. Motivated by a recent work of Setiono, a path-following algorithm for linear programming using both logarithmic and quadratic penalty functions is proposed. In the algorithm, a logarithmic and a quadratic penalty is placed on, respectively, the nonnegativity constraints and an arbitrary subset of the equality constraints; Newton's method is applied to solve the penalized problem, and after each Newton step the penalty parameters are decreased. This algorithm maintains neither primal nor dual feasibility and does not require a Phase I. It is shown that if the initial iterate is chosen appropriately and the penalty parameters are decreased to zero in a particular way, then the algorithm is linearly convergent. Numerical results are also presented suggesting that the algorithm may be competitive with interior point algorithms in practice, requiring typically between 30–45 iterations to accurately solve each Netlib problem tested.

Key words. linear program, path-following, Newton step, penalty function

AMS subject classifications. 49, 90

1. Introduction. Since the pioneering work of Karmarkar [Kar84], much interest has focused on solving linear programs using interior point algorithms. These interior point algorithms may be classified roughly as either (i) projective-scaling (or potential reduction), (ii) affine-scaling, or (iii) path-following. We will not attempt to review the literature on this subject, which is vast (see for example [Meg89], [Tod89] for surveys). Our interest is in algorithms of the path-following type, of the sort discussed in [GaZ81]. These interior point algorithms typically penalize the nonnegativity constraints by a logarithmic function and use Newton's method to solve the penalized problem, with the penalty parameters decreased after each Newton step (see, for example, [Gon89], [GMSTW86], [KMY89], [MoA87], [Ren88], [Tse89]).

One disadvantage of interior point algorithms is the need for an initial interior feasible solution. A common technique for handling this is to add an artificial column (see [AKRV89], [BDDW89], [GMSTW86], [Lus90], [MMS89], [MSSPB89], [MoM87]), but this itself has disadvantages. For example, the cost of the artificial column must be estimated, and some type of rank-1 updating is needed to solve each least square problem which can significantly increase the solution time and degrade the numerical accuracy of the solutions.

Recently, Setiono [Set89] proposed an interesting algorithm that combines features of a path-following algorithm with those of the method of multipliers [HaB70], [Hes69], [Pow69] (also see [Roc76], [Ber82]). This algorithm does not require a feasible solution to start and is comparable to interior point algorithms both in terms of work per iteration and, according to the numerical results reported in [Set89], in terms of the total number of iterations. To describe the basic idea in Setiono's algorithm, consider

* Received by the editors April 16, 1990; accepted for publication (in revised form) June 11, 1992. This research was supported by the U.S. Army Research Office, contract DAAL03-86-K-0171, and was conducted while the author was with the Laboratory for Information and Decision Systems and the Center for Intelligent Control Systems, Massachusetts Institute of Technology, Cambridge.

† Department of Mathematics, GN-50, University of Washington, Seattle, Washington 98195 (tseng@math.washington.edu).

a linear program in the standard dual form

$$\begin{array}{ll} \text{minimize} & -b^T p \\ \text{subject to} & t + A^T p = c, \quad t \geq 0, \end{array}$$

where A is some matrix and b and c are vectors of appropriate dimension. Let us attach a Lagrange multiplier vector x to the constraints $t + A^T p = c$ and apply the method of multipliers to the above linear program. This produces the following iterations

$$x^{k+1} = x^k + \frac{1}{\epsilon^k}(t^k + A^T p^k - c), \quad k = 1, 2, \dots,$$

where $\{\epsilon^k\}$ is a sequence of monotonically decreasing positive scalars and (t^k, p^k) is some (inexact) solution of the augmented Lagrangian subproblem

$$(1.1) \quad \begin{array}{ll} \text{minimize} & -b^T p + (x^k)^T(t + A^T p - c) + \frac{1}{2\epsilon^k} \|t + A^T p - c\|^2 \\ \text{subject to} & t \geq 0. \end{array}$$

(An advantage of the above multiplier iterations is that they do not need a feasible solution to start.) A key issue associated with the above multiplier iterations concerns the efficient generation of an inexact solution (t^k, p^k) of the convex quadratic program (1.1) for each k . (Note that as ϵ^k decreases, the objective function of (1.1) becomes progressively more ill-conditioned.) Setiono's algorithm may be viewed as the method of multipliers in which (t^k, p^k) is generated according to the following scheme, reminiscent of the path-following idea: Add a logarithmic penalty function $-\gamma^k \sum_{j=1}^m \ln(t_j)$ to the objective of (1.1), where γ^k is some positive scalar monotonically decreasing with k , and apply a single Newton step, starting from (t^{k-1}, p^{k-1}) , to the resulting problem. (If the t^k thus obtained lies outside the positive orthant, it is moved back towards t^{k-1} until it becomes positive.¹)

In this paper, inspired by the preceding work of Setiono, we study an algorithm that also adds to the objective a quadratic penalty on the equality constraints and a logarithmic penalty on the nonnegativity constraints; and then solves the penalized problem using Newton's method, with the penalty parameters decreased after each Newton step. Unlike Setiono's algorithm, our algorithm does not use the multiplier vector x^k (so it may be viewed as a pure penalty method) and allows any subset of the equality constraints to be penalized. We show that if the problem is primal non-degenerate and the iterates start near the optimal solution of the initial penalized problem, then the penalty parameters can be decreased at the rate of a geometric progression and the iterates converge linearly. To the best of our knowledge, this is the first (global) linear convergence result for an noninterior point path-following algorithm. We also present numerical results indicating that the algorithm may potentially be competitive with interior point algorithms in practice. We remark that penalty methods that use either the quadratic or the logarithmic penalty function have been well studied (see, for example, [Ber82], [FiM68], [Fri57], [JiO78], [Man84],

¹ More precisely, t^k is given by the formula

$$t^k = t^{k-1} + .98\lambda^k \Delta t^k,$$

where Δt^k is the Newton direction (projected onto the space of t) and λ^k is the largest $\lambda \in (0, 1]$ for which $t^{k-1} + \lambda \Delta t^k$ is nonnegative. (The choice of .98 is arbitrary—any number between 0 and 1 would do.)

[WBD88]), but very little is known about penalty methods that use both types of penalty functions (called mixed interior point-exterior point algorithms in [FiM68]).

Upon completion of this paper, the author learned from Professor O. L. Mangasarian that the algorithm discussed here is essentially the IDLN (Interior Dual Least Norm) algorithm described in the recent Ph.D. thesis of Setiono [Set90] (also see the report [Set91]), with minor differences in the choice of the penalty parameters. The same reference includes an (asymptotic) linear rate convergence analysis and extensive numerical results showing that the IDLN algorithm outperforms the popular simplex code MINOS by a factor of 2 or more on the 63 Netlib problems tested. In short, Setiono's independent work provides further evidence of the practical efficiency of the mixed interior point-exterior point solution approach. Finally, we note that while this paper was under review, other noninterior point methods relating to that studied here have been proposed. One method, brought to our attention by one of the referees, is a certain augmented Lagrangian algorithm for stochastic programming (see [MuR90]); another method is a primal-dual exterior point algorithm for linear programming (see [KMM91]). However, neither of these methods has been shown to possess the nice theoretical/numerical properties enjoyed by the algorithm studied here. For example, no convergence result is given for the method in [MuR90] and no numerical or rate of convergence result is given for the method in [KMM91].

This paper proceeds as follows: In §2 we describe the basic algorithm; in §3 we analyze its convergence; and in §4 we recount our numerical experience with it. In §5 we discuss extensions of this work.

In our notation, every vector is a column vector in some k -dimensional real space \Re^k , and superscript T denotes transpose. For any vector x , we denote by x_j the j th coordinate of x , by $\text{Diag}(x)$ the diagonal matrix whose j th diagonal entry is x_j , and by $\|x\|_1$, $\|x\|$, $\|x\|_\infty$ the L_1 -norm, the L_2 -norm, and the L_∞ -norm of x , respectively. For any matrix A , we denote by A_j the j th column of A . We also denote by e the vector of 1's (whose dimension will be clear from the context) and denote by $\ln(\cdot)$ the natural logarithm function.

2. Algorithm description. Let A be an $n \times m$ matrix, B be an $l \times m$ matrix, b be an n -vector, c be an m -vector, and d be an l -vector. Consider the following linear program associated with A , B , b , c , and d :

$$\begin{aligned}
 (\mathcal{D}) \quad & \text{minimize } -b^T p \\
 & \text{subject to } t + A^T p = c, \quad Bt = d, \quad t \geq 0,
 \end{aligned}$$

which we call the dual problem. The dual problem may be viewed as a standard linear program in t , in which we arbitrarily partition the equality constraints into two subsets and express one subset in the generator form $t + A^T p = c$. (To see this, consider the problem of minimizing $a^T t$ subject to $t \in R \cap S$, $t \geq 0$, where a is an m -vector and R and S are affine sets in \Re^m . We can always express $R = \{ t \mid t = c - A^T p \text{ for some } p \}$ and $S = \{ t \mid Bt = d \}$ for some matrices A and B and some vectors c and d . Using $t = c - A^T p$ to eliminate t from the objective function, the problem is now in the form (\mathcal{D}) .) The constraints $t + A^T p = c$ can be thought of as the complicating constraints which, if removed, would make (\mathcal{D}) much easier to solve. The advantages of splitting the constraints in this manner will be explained at the end of §2. Finally, we note that the form in which we write the equality constraints is unimportant, and the above form is adopted for notational convenience only.

By attaching Lagrange multiplier vectors x and y to the constraints $c - A^T p = t$

and $Bt = d$, respectively, we obtain the following dual of (\mathcal{D}) :

$$(\mathcal{P}) \quad \begin{aligned} & \text{minimize } c^T x + d^T y \\ & \text{subject to } Ax = b, \quad x + B^T y \geq 0, \end{aligned}$$

which we call the primal problem.

We make the following blanket assumptions, which are standard for interior point algorithms, regarding (\mathcal{D}) and (\mathcal{P}) .

Assumption A.

- (a) $\{ (x, y) \mid Ax = b, x + B^T y > 0 \}$ is nonempty.
- (b) $\{ (t, p) \mid Bt = d, t > 0, t + A^T p = c \}$ is nonempty.
- (c) A has full row rank.

It is well known that, under parts (a) and (b) of Assumption A, both (\mathcal{D}) and (\mathcal{P}) have nonempty bounded optimal solution sets.

Consider the dual problem (\mathcal{D}) . Suppose that we place a quadratic penalty on the constraints $t + A^T p = c$ with a penalty parameter $1/\epsilon$ ($\epsilon > 0$) and we place a logarithmic penalty on the constraints $t \geq 0$ with a penalty parameter $\gamma > 0$. This gives the following approximation to (\mathcal{D}) :

$$(\mathcal{D}_{\epsilon, \gamma}) \quad \begin{aligned} & \text{minimize } f_{\epsilon, \gamma}(t, p) \\ & \text{subject to } Bt = d, \quad t > 0, \end{aligned}$$

where $f_{\epsilon, \gamma} : (0, \infty)^m \times \mathbb{R}^n$ is the penalized objective function given by

$$(2.1) \quad f_{\epsilon, \gamma}(t, p) = \frac{1}{2} \|c - t - A^T p\|^2 - \epsilon \gamma \sum_{j=1}^m \ln(t_j) - \epsilon b^T p \quad \forall t > 0, \forall p.$$

The penalized problem $(\mathcal{D}_{\epsilon, \gamma})$ has the advantage that its objective function $f_{\epsilon, \gamma}$ is twice differentiable and the Hessian $\nabla^2 f_{\epsilon, \gamma}$ is positive definite everywhere (via Assumption A(c)). Since the feasible set of $(\mathcal{D}_{\epsilon, \gamma})$ is nonempty (by Assumption A(b)) and its intersection with any level set of $f_{\epsilon, \gamma}$ is bounded (by Assumption A(a)), it is readily seen that $(\mathcal{D}_{\epsilon, \gamma})$ has an optimal solution which, by the strict convexity of $f_{\epsilon, \gamma}$, is unique.

Note 1. We can use penalty functions other than the quadratic and the logarithmic. For example, we can use a cubic in place of the quadratic and $t_j \ln(t_j)$ in place of $-\ln(t_j)$. The quadratic and the logarithmic function, however, have nice properties (such as the second derivative of the logarithmic function equals minus the square of its first derivative) which make global convergence analysis possible.

It is well known (see [Roc70]) that (t, p) is the optimal solution of $(\mathcal{D}_{\epsilon, \gamma})$ if and only if it satisfies, together with some $u \in \mathbb{R}^l$, the Kuhn–Tucker conditions

$$(2.2) \quad t > 0, \quad Bt = d, \quad \nabla f_{\epsilon, \gamma}(t, p) + \begin{pmatrix} B^T u \\ 0 \end{pmatrix} = 0.$$

Straightforward calculation using (2.1) finds that

$$(2.3) \quad \nabla f_{\epsilon, \gamma}(t, p) = \begin{pmatrix} t + A^T p - c - \epsilon \gamma (T)^{-1} e \\ A(t + A^T p - c) - \epsilon b \end{pmatrix},$$

$$(2.4) \quad \nabla^2 f_{\epsilon, \gamma}(t, p) = \begin{pmatrix} I + \epsilon \gamma (T)^{-2} & A^T \\ A & AA^T \end{pmatrix},$$

where $T = \text{Diag}(t)$. The above formulas will be used extensively in the subsequent analysis. Note that $\nabla^2 f_{\epsilon, \gamma}$ is ill-conditioned at the boundary of its domain.

It is not difficult to show that, as ϵ and γ tend to zero, the optimal solution of $(\mathcal{D}_{\epsilon, \gamma})$ approaches the optimal solution set of (\mathcal{D}) (see Lemma 3.1). This suggests the following algorithm for solving $(\mathcal{D}_{\epsilon, \gamma})$. At each iteration, we are given ϵ , γ and a (t, p) which is an approximate solution of $(\mathcal{D}_{\epsilon, \gamma})$; we apply a Newton step to $(\mathcal{D}_{\epsilon, \gamma})$ at (t, p) to generate a new (t, p) and then we decrease ϵ and γ . In other words, we consider a sequence of penalized problems $(\mathcal{D}_{\epsilon^k, \gamma^k})$, $k = 1, 2, \dots$, with $\epsilon^k \downarrow 0$ and $\gamma^k \downarrow 0$, and we use a Newton step to follow the optimal solution of one penalized problem to that of the next. We now formally state this algorithm, which we call the QLPPF (short for Quadratic-Logarithmic Penalty Path-Following) algorithm.

QLPPF Algorithm

Iteration 0. Choose $\epsilon^1 > 0$ and $\gamma^1 > 0$. Choose $(t^1, p^1) \in (0, \infty)^m \times \mathbb{R}^n$ with $Bt^1 = d$.

Iteration k . Given $(t^k, p^k) \in (0, \infty)^m \times \mathbb{R}^n$ with $Bt^k = d$, compute $(\Delta t^k, \Delta p^k, u^k)$ to be a solution of

$$(2.5) \quad \nabla^2 f_{\epsilon^k, \gamma^k}(t^k, p^k) \begin{pmatrix} \Delta t^k \\ \Delta p^k \end{pmatrix} + \nabla f_{\epsilon^k, \gamma^k}(t^k, p^k) + \begin{pmatrix} B^T u^k \\ 0 \end{pmatrix} = 0, \quad B\Delta t^k = 0,$$

and set

$$(2.6) \quad t^{k+1} = t^k + \Delta t^k, \quad p^{k+1} = p^k + \Delta p^k,$$

$$(2.7) \quad \gamma^{k+1} = \alpha^k \gamma^k, \quad \epsilon^{k+1} = \alpha^k \epsilon^k,$$

where α^k is some scalar in $(0, 1)$.

Note 2. As we noted earlier, the QLPPF Algorithm is closely linked to the algorithms proposed by Setiono. In particular, it can be seen from (2.3), (2.4) that, in the special case where B is the zero matrix, the direction finding problem (2.5) is identical to that in the IDLN algorithm of Setiono (see [Set91, Eq. (16)]) and differs from that in the IDPP algorithm of Setiono (see [Set89, Eq. (6)]) by only an order ϵ^k term on the right-hand side (which tends to zero as ϵ^k tends to zero).

Note 3. A unique feature of the QLPPF Algorithm lies in its handling of the two sets of constraints $t + A^T p = c$ and $Bt = d$, whereby quadratic penalties are placed only on the first set while the second set is maintained to be satisfied at all iterations. This feature has the advantage that it enables the QLPPF Algorithm to provide a unified framework for interior point methods and exterior point methods. As an example, if we put all the equality constraints into $Bt = d$ (and correspondingly set $A = I$ and $c = 0$), then it can be seen that the QLPPF Algorithm with $\gamma^k = 1$ for all k reduces to the well-known primal path-following algorithm for maximizing $b^T t$ subject to the constraints $Bt = d$, $t \geq 0$ (see [Gon89], [Tse89], [Ye87]). If we put all the equality constraints into $t + A^T p = c$ (and correspondingly set $B = 0$ and $d = 0$), then, as was noted above, the QLPPF Algorithm reduces to the IDLN algorithm for maximizing $b^T p$ subject to the constraints $t + A^T p = c$, $t \geq 0$. For other choices of constraint splitting, we obtain algorithms somewhere in between. We can then envision choosing a constraint splitting so the corresponding QLPPF Algorithm is in some sense most efficient (e.g., fastest convergence) for the given problem. Alternatively, we may choose a constraint splitting so to ensure that certain “critical” equality constraints are satisfied exactly at all iterations (by putting these constraints into $Bt = d$).

3. Global convergence. In this section, we show that if (\mathcal{D}) is in some sense primal nondegenerate and if (t^1, p^1) is “close” to the optimal solution of $(\mathcal{D}_{\epsilon^1, \gamma^1})$ in the QLPPF Algorithm, then, by decreasing the α^k 's at an appropriate rate, the iterates $\{(t^k, p^k)\}$ generated by the QLPPF Algorithm approach the optimal solution set of (\mathcal{D}) (see Theorem 3.4). Because the Hessian $\nabla^2 f_{\epsilon, \gamma}$ is ill-conditioned at the boundary of its domain, the proof of this result is quite involved and relies critically on finding a suitable Lyapunov (i.e., merit) function to monitor the progress of the algorithm.

For any $\epsilon > 0$ and $\gamma > 0$, let $\rho_{\epsilon, \gamma} : (0, \infty)^m \times \mathbb{R}^n \times \mathbb{R}^l$ be the function given by

$$(3.1) \quad \rho_{\epsilon, \gamma}(t, p, u) = \max \left\{ \frac{1}{\epsilon \gamma} \|T(c - t - A^T p - B^T u) + \epsilon \gamma e\|, \right. \\ \left. \frac{1}{\sqrt{\epsilon \gamma}} \|A(c - t - A^T p) + \epsilon b\| \right\} \quad \forall t > 0, \forall p, \forall u,$$

where $T = \text{Diag}(t)$. From (2.2) and (2.3) we see that (t, p) is a solution of $(\mathcal{D}_{\epsilon, \gamma})$ if and only if (t, p) satisfies $Bt = d$, $t > 0$ and $\rho_{\epsilon, \gamma}(t, p, u) = 0$ for some u . Hence $\rho_{\epsilon, \gamma}$ acts as a Lyapunov function which measures how far (t, p) is from solving $(\mathcal{D}_{\epsilon, \gamma})$. This notion is made precise in the following lemma.

For any $\epsilon > 0$, let

$$(3.2) \quad \mathcal{Y}_\epsilon = \{ (t, p) \mid A(c - t - A^T p) = -\epsilon b, \quad Bt = d, \quad t > 0 \}.$$

The following lemma shows that any $(t, p) \in \mathcal{Y}_\epsilon$ that satisfies $\rho_{\epsilon, \gamma}(t, p, u) \leq \beta$, for some u , is within an order of $\epsilon + (1 + \beta)\gamma$ of being an optimal solution of (\mathcal{D}) .

LEMMA 3.1. *Fix any $\epsilon > 0$, $\gamma > 0$ and $\beta \in (0, 1]$. For any $(t, p) \in \mathcal{Y}_\epsilon$ that satisfies $\rho_{\epsilon, \gamma}(t, p, u) \leq \beta$ for some u , the vector (x, y) given by*

$$(3.3) \quad x = (t + A^T p - c)/\epsilon, \quad y = u/\epsilon$$

is feasible for (\mathcal{P}) and, together with (t, p) , satisfies

$$\begin{aligned} c^T x + d^T y &\leq v^* + \eta^* \epsilon + m(1 + \beta)\gamma, \\ b^T p &\geq v^* - m(1 + \beta)\gamma, \\ \|t + A^T p - c\|^2 &\leq 2(\eta^*(\epsilon)^2 + m(1 + \beta)\epsilon\gamma), \end{aligned}$$

where v^ denotes the optimal cost of (\mathcal{P}) and $\eta^* = \min\{ \|x^*\|^2/2 \mid (x^*, y^*) \text{ is an optimal solution of } (\mathcal{P}) \text{ for some } y^* \}$.*

Proof. Since (t, p) is in \mathcal{Y}_ϵ , it follows from the definition of \mathcal{Y}_ϵ (see (3.2)) and (3.3) that $Ax = b$. Since $\rho_{\epsilon, \gamma}(t, p, u) \leq \beta$, we have from the definition of $\rho_{\epsilon, \gamma}$ (see (3.1)) and (3.3) that $\|T(-x - B^T y) + \gamma e\| \leq \gamma\beta$, where $T = \text{Diag}(t)$. Thus

$$(3.4) \quad \gamma(1 - \beta)e \leq T(x + B^T y) \leq \gamma(1 + \beta)e.$$

Since $t > 0$ and $\beta \leq 1$, the first inequality in (3.4) yields $x + B^T y \geq \gamma(1 - \beta)T^{-1}e \geq 0$. Hence, (x, y) is feasible for (\mathcal{P}) .

Let

$$\begin{aligned} d_\epsilon(\tau, \pi) &= b^T \pi - \frac{1}{2\epsilon} \|\tau + A^T \pi - c\|^2 \quad \forall (\tau, \pi), \\ p_\epsilon(\xi, \psi) &= \frac{\epsilon}{2} \|\xi\|^2 + c^T \xi + d^T \psi \quad \forall (\xi, \psi). \end{aligned}$$

Straightforward algebra using (3.3) yields

$$p_\epsilon(x, y) = d_\epsilon(t, p) + t^T(x + B^T y).$$

Also, it follows from strong duality for a convex quadratic program and its dual that

$$\min_{(\xi, \psi)} \{ p_\epsilon(\xi, \psi) \mid A\xi = b, \xi + B^T \psi \geq 0 \} = \max_{(\tau, \pi)} \{ d_\epsilon(\tau, \pi) \mid B\tau = d, \tau \geq 0 \}.$$

Thus, by letting (x^*, y^*) be any optimal solution of (\mathcal{P}) with $\|x^*\|^2/2 = \eta^*$, we obtain from the above two relations and $c^T x^* + d^T y^* = v^*$ that

$$\begin{aligned} \frac{\epsilon}{2} \|x\|^2 + c^T x + d^T y &= p_\epsilon(x, y) \\ &= d_\epsilon(t, p) + t^T(x + B^T y) \\ &\leq p_\epsilon(x^*, y^*) + t^T(x + B^T y) \\ &= \epsilon \eta^* + v^* + t^T(x + B^T y), \end{aligned}$$

Similarly, by letting (t^*, p^*) be any optimal solution of (\mathcal{D}) , we obtain from the same two relations and the facts $b^T p^* = v^*$ and $t^* + A^T p^* = c$ that

$$\begin{aligned} b^T p - \frac{1}{2\epsilon} \|t + A^T p - c\|^2 &= d_\epsilon(t, p) \\ &= p_\epsilon(x, y) - t^T(x + B^T y) \\ &\geq d_\epsilon(t^*, p^*) - t^T(x + B^T y) \\ &= v^* - t^T(x + B^T y). \end{aligned}$$

Since $0 \leq t^T(x + B^T y) \leq m(1 + \beta)\gamma$ (cf. (3.4)), the above two relations yield

$$\begin{aligned} c^T x + d^T y &\leq v^* + \epsilon \eta^* + m(1 + \beta)\gamma \\ b^T p &\geq v^* - m(1 + \beta)\gamma. \end{aligned}$$

Finally, since (x, y) is feasible for (\mathcal{P}) so that $c^T x + d^T y \geq v^*$, the first of these two relations also yields

$$\frac{\epsilon}{2} \|x\|^2 + v^* \leq \epsilon \eta^* + v^* + m(1 + \beta)\gamma.$$

Canceling v^* from both sides and using (3.3) completes the proof. \square

Since we are dealing with linear programs, Lemma 3.1 implies that, as $\epsilon \downarrow 0$ and $\gamma \downarrow 0$, the (x, y) given by (3.3) approaches the optimal solution set of (\mathcal{P}) and (t, p) approaches the optimal solution set of (\mathcal{D}) . In fact, it suffices to decrease ϵ and γ as far as $2^{\kappa L}$, where κ is some scalar constant and L is the size of the problem encoding in binary (defined as, say, in [Kar84]), at which time an optimal solution of (\mathcal{P}) and of (\mathcal{D}) can be recovered by using the techniques described in, for example, [Kar84] and [PaS82].

For each $\lambda > 0$, let $\theta_\lambda : (0, \infty)^m \rightarrow [0, \infty)$ be the function given by

$$(3.5) \quad \theta_\lambda(t) = (\|E + ED^{1/2}A^T FAD^{1/2}E\| + \|ED^{1/2}A^T F\|)^2 \quad \forall t > 0,$$

where

$$(3.6) \quad D = (I + \lambda T^{-2})^{-1},$$

$$(3.7) \quad E = I - D^{1/2}B^T[BDB^T]^{-1}BD^{1/2},$$

$$(3.8) \quad F = [A(I - D^{1/2}ED^{1/2})A^T]^{-1},$$

and $T = \text{Diag}(t)$. (F is well defined because $\|E\| \leq 1$ (E is a projection matrix) and $\|D\| < 1$, so that $I - D^{1/2}ED^{1/2}$ is positive definite. We also use the assumption that A has full row rank.) The quantity $\theta_\lambda(t)$ estimates the norm squared of certain projection-like operator depending on λ and t , and it will be used extensively in our analysis. In general, $\theta_\lambda(t)$ is rather cumbersome to evaluate, but, as we shall see, it suffices for our analysis to bound $\theta_\lambda(t)$ from above (see Lemma 3.3(b)).

3.1. Analyzing a Newton step. In this subsection we prove a key lemma that states that if (\bar{t}, \bar{p}) is “close” to the optimal solution of $(\mathcal{D}_{\bar{\epsilon}, \bar{\gamma}})$, then (t, p) generated by applying one Newton step to (\mathcal{D}) at (\bar{t}, \bar{p}) is close to the optimal solution of $(\mathcal{D}_{\epsilon, \gamma})$ for some $\epsilon < \bar{\epsilon}$ and some $\gamma < \bar{\gamma}$. The notion of “closeness” is measured by the Lyapunov function $\rho_{\epsilon, \gamma}$ and the proof of the lemma is based on the ideas used in [Tse89, §2].

LEMMA 3.2. *For any $\bar{\epsilon} > 0$, any $\bar{\gamma} > 0$ and any $(\bar{t}, \bar{p}, \bar{u}) \in (0, \infty)^m \times \mathbb{R}^n \times \mathbb{R}^l$ with $B\bar{t} = d$, let (t, p, u) be given by*

$$(3.9) \quad t = \bar{t} + \Delta t,$$

$$(3.10) \quad p = \bar{p} + \Delta p,$$

where u and $(\Delta t, \Delta p)$ together solve the following system of linear equations

$$(3.11) \quad \nabla^2 f_{\bar{\epsilon}, \bar{\gamma}}(\bar{t}, \bar{p}) \begin{pmatrix} \Delta t \\ \Delta p \end{pmatrix} + \nabla f_{\bar{\epsilon}, \bar{\gamma}}(\bar{t}, \bar{p}) + \begin{pmatrix} B^T u \\ 0 \end{pmatrix} = 0, \quad B\Delta t = 0.$$

Suppose that $\rho_{\bar{\epsilon}, \bar{\gamma}}(\bar{t}, \bar{p}, \bar{u}) \leq \beta$ for some $\beta < \min\{1, 1/\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})\}$. Then the following hold:

(a) $(t, p) \in \mathcal{Y}_{\bar{\epsilon}}$.

(b) For any α satisfying

$$(3.12) \quad \max \left\{ \sqrt{\frac{\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})(\beta)^2 + \sqrt{m}}{\beta + \sqrt{m}}}, \frac{1}{1 + \beta\sqrt{\bar{\gamma}/\bar{\epsilon}}/\|b\|} \right\} \leq \alpha \leq 1,$$

we have $\rho_{\alpha\bar{\epsilon}, \alpha\bar{\gamma}}(t, p, u) \leq \beta$.

Proof. Let

$$(3.13) \quad \bar{r} = \bar{T}(c - \bar{t} - A^T \bar{p} - B^T \bar{u}) + \bar{\epsilon}\bar{\gamma}e,$$

$$(3.14) \quad \bar{s} = A(c - \bar{t} - A^T \bar{p}) + \bar{\epsilon}b,$$

where $\bar{T} = \text{Diag}(\bar{t})$. Then, by (3.1),

$$(3.15) \quad \max\{\|\bar{r}\|/(\bar{\epsilon}\bar{\gamma}), \|\bar{s}\|/\sqrt{\bar{\epsilon}\bar{\gamma}}\} = \rho_{\bar{\epsilon}, \bar{\gamma}}(\bar{t}, \bar{p}, \bar{u}) \leq \beta.$$

By using (2.3), (2.4) and (3.13), (3.14), we write (3.11) equivalently as

$$\begin{aligned} D^{-1}\Delta t + A^T \Delta p + B^T(u - \bar{u}) &= \bar{T}^{-1}\bar{r}, \\ A\Delta t + AA^T \Delta p &= \bar{s}, \\ B\Delta t &= 0, \end{aligned}$$

where for convenience we let $D = (I + \bar{\epsilon}\bar{\gamma}\bar{T}^{-2})^{-1}$. Solving for Δt gives

$$\Delta t = D^{1/2}(E + ED^{1/2}A^T F AD^{1/2}E)D^{1/2}\bar{T}^{-1}\bar{r} - D^{1/2}ED^{1/2}A^T F \bar{s},$$

where we let $E = I - D^{1/2}B^T[BDB^T]^{-1}BD^{1/2}$ and $F = [A(I - D^{1/2}ED^{1/2})A^T]^{-1}$. (F is well defined by the same reasoning that the matrix given by (3.8) is well defined.) Then, we can bound $\bar{T}^{-1}\Delta t$ as follows:

$$\|\bar{T}^{-1}\Delta t\| \leq \|\bar{T}^{-1}D^{1/2}\|^2 \|E + ED^{1/2}A^T F AD^{1/2}E\| \|\bar{r}\| + \|\bar{T}^{-1}D^{1/2}\| \|ED^{1/2}A^T F\| \|\bar{s}\|.$$

Since $\bar{T}^{-2}D$ is diagonal and each of its diagonal entry is less than $1/(\bar{\epsilon}\bar{\gamma})$, we obtain that $\|\bar{T}^{-2}D\| < 1/(\bar{\epsilon}\bar{\gamma})$ and hence

$$\begin{aligned} \|\bar{T}^{-1}\Delta t\| &\leq \|E + ED^{1/2}A^T FAD^{1/2}E\| \|\bar{r}\|/(\bar{\epsilon}\bar{\gamma}) + \|ED^{1/2}A^T F\| \|\bar{s}\|/\sqrt{\bar{\epsilon}\bar{\gamma}} \\ (3.16) \quad &\leq \sqrt{\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})}\beta, \end{aligned}$$

where the last inequality follows from (3.15) and the definition of \bar{T} , D , E , F , and $\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})$ (see (3.5)–(3.8)).

Now, by using (2.3) and (2.4), we can write (3.11) equivalently as $B\Delta t = 0$ and

$$(3.17) \quad \bar{\epsilon}\bar{\gamma}\bar{T}^{-1}\Delta t = \bar{T}(c - \bar{t} - \Delta t - A^T\bar{p} - A^T\Delta p - B^Tu) + \bar{\epsilon}\bar{\gamma}e,$$

$$(3.18) \quad 0 = A(c - \bar{t} - \Delta t - A^T\bar{p} - A^T\Delta p) + \bar{\epsilon}b,$$

so from (3.9) and (3.10) we obtain

$$\begin{aligned} T(c - t - A^Tp - B^Tu) + \bar{\epsilon}\bar{\gamma}e &= (\bar{T} + \Delta T)(c - \bar{t} - \Delta t - A^T\bar{p} - A^T\Delta p - B^Tu) + \bar{\epsilon}\bar{\gamma}e \\ &= \bar{T}(c - \bar{t} - \Delta t - A^T\bar{p} - A^T\Delta p - B^Tu) \\ &\quad + \bar{\epsilon}\bar{\gamma}e + \Delta T(c - \bar{t} - \Delta t - A^T\bar{p} - A^T\Delta p - B^Tu) \\ &= \bar{\epsilon}\bar{\gamma}\bar{T}^{-1}\Delta t + \Delta T(c - \bar{t} - \Delta t - A^T\bar{p} - A^T\Delta p - B^Tu) \\ &= \bar{T}^{-1}\Delta T(\bar{\epsilon}\bar{\gamma}e) + \bar{T}(c - \bar{t} - \Delta t - A^T\bar{p} - A^T\Delta p - B^Tu) \\ &= \bar{\epsilon}\bar{\gamma}\bar{T}^{-2}\Delta T\Delta t, \end{aligned}$$

where $T = \text{Diag}(t)$, $\Delta T = \text{Diag}(\Delta t)$, and the third and the last equality follow from (3.17). This implies

$$\begin{aligned} \|T(c - t - A^Tp - B^Tu) + \bar{\epsilon}\bar{\gamma}e\| &\leq \bar{\epsilon}\bar{\gamma}\|\bar{T}^{-2}\Delta T\Delta t\| \\ &\leq \bar{\epsilon}\bar{\gamma}\|\bar{T}^{-2}\Delta T\Delta t\|_1 \\ &= \bar{\epsilon}\bar{\gamma}\|\bar{T}^{-1}\Delta t\|^2 \\ (3.19) \quad &\leq \bar{\epsilon}\bar{\gamma}\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})(\beta)^2, \end{aligned}$$

where the third inequality follows from (3.16).

(a) We have from (3.16) and the hypothesis $\beta < 1/\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})$ that $\|\bar{T}^{-1}\Delta t\|^2 \leq \theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})(\beta)^2 < \beta < 1$, so (3.9) and $\bar{t} > 0$ yields $t = \bar{t} + \Delta t > 0$. Also, $B\bar{t} = d$ together with $B\Delta t = 0$ (see (3.11)) and (3.9) yields $Bt = d$, and (3.18) together with (3.9), (3.10) yields $0 = A(c - t - A^Tp) + \bar{\epsilon}b$. Hence $(t, p) \in \mathcal{Y}_{\bar{\epsilon}}$.

(b) Fix any α satisfying (3.12) and let $\gamma = \alpha\bar{\gamma}$, $\epsilon = \alpha\bar{\epsilon}$. (Note that because $\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})\beta < 1$, the left-hand quantity in (3.12) is strictly less than 1, so such an α exists.) Let

$$r = T(c - t - A^Tp - B^Tu) + \epsilon\gamma e.$$

Then the triangle inequality and (3.19) imply

$$\begin{aligned} \|r\|/(\epsilon\gamma) &\leq \|T(c - t - A^Tp - B^Tu) + \bar{\epsilon}\bar{\gamma}e\|/(\epsilon\gamma) + (1 - (\alpha)^2)\sqrt{m}/(\alpha)^2 \\ &\leq \theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})(\beta)^2/(\alpha)^2 + (1/(\alpha)^2 - 1)\sqrt{m}, \end{aligned}$$

which together with the fact (see (3.12)) $(\theta_{\bar{\epsilon}\bar{\gamma}}(\bar{t})(\beta)^2 + \sqrt{m})/(\beta + \sqrt{m}) \leq (\alpha)^2$, yields

$$(3.20) \quad \|r\|/(\epsilon\gamma) \leq \beta.$$

Let $s = A(c - t - A^T p) + \epsilon b$. By using (3.9), (3.10), and (3.18), we have

$$s = A(c - \bar{t} - \Delta t - A^T \bar{p} - A^T \Delta p) + \alpha \bar{\epsilon} b = (\alpha - 1) \bar{\epsilon} b,$$

which together with the fact (see (3.12))

$$1/(1 + \beta\sqrt{\bar{\gamma}/\bar{\epsilon}}/\|b\|) \leq \alpha \leq 1,$$

yields

$$\|s\|/\sqrt{\epsilon\gamma} = (1/\alpha - 1)\|b\|\sqrt{\bar{\epsilon}/\bar{\gamma}} \leq \beta.$$

This together with (3.20) and the definition of $\rho_{\epsilon,\gamma}(t, p, u)$ (see (3.1)) proves our claim. \square

3.2. Bounds. Lemma 3.2 shows that if the rate α at which the penalty parameters ϵ and γ are decreased is not too small (see (3.12)), then a single Newton step suffices to keep the current iterate close to the optimal solution of the penalized problem $(\mathcal{D}_{\epsilon,\gamma})$. Thus, in order to establish the (linear) convergence of the QLPPF Algorithm, it suffices to bound α away from 1 which, according to (3.12), amounts to bounding $\theta_{\epsilon,\gamma}(t)$ by some quantity independent of t and ϵ, γ . It is not difficult to see that such a bound does not exist for arbitrary t . Fortunately, we need to consider only those t that, together with some p , are close to the optimal solution of $(\mathcal{D}_{\epsilon,\gamma})$, in which case, as we show below, such a bound does exist (provided that a certain primal nondegeneracy assumption also holds). The proof of this is somewhat intricate: For ϵ and γ large, we argue by showing that t cannot be too large, i.e., of the order $\epsilon + \gamma + 1$ (see Lemma 3.3(a)) and, for ϵ and γ small, we argue by showing that, under the primal nondegeneracy assumption, the columns of A corresponding to those components of t which are small (i.e., of the order γ) are of rank n .

LEMMA 3.3. (a) *For all $\epsilon > 0$, all $\gamma > 0$, and all $(t, p) \in \mathcal{Y}_\epsilon$ satisfying $\rho_{\epsilon,\gamma}(t, p, u) \leq 1$ for some u , we have*

$$\|t\| \leq M_1(\epsilon + \gamma + 1),$$

where $M_1 > 0$ is some scalar depending on A, B, b, c , and d only.

(b) *Suppose that (\mathcal{P}) is primal nondegenerate in the sense that, for every optimal solution (x^*, y^*) of (\mathcal{P}) , those columns of A corresponding to the positive components of $x^* + B^T y^*$ have rank n . Then, for all $\epsilon > 0$, all $\gamma > 0$, all $\lambda \geq \epsilon\gamma/2$, and all $(t, p) \in \mathcal{Y}_\epsilon$ satisfying $\rho_{\epsilon,\gamma}(t, p, u) \leq 1$ for some u , we have*

$$\theta_\lambda(t) \leq \psi(\epsilon/\gamma),$$

where we define

$$(3.21) \quad \psi(\omega) = M_2(\sqrt{\omega} + 1/\sqrt{\omega})^4 \quad \forall \omega > 0,$$

with $M_2 \geq 1$ some scalar depending on A, B, b, c , and d only.

Proof. (a) The proof is by contradiction. Suppose the contrary, so that there exists a sequence $\{(t^k, p^k, u^k, \epsilon^k, \gamma^k)\}$ such that $\epsilon^k > 0$ and $\gamma^k > 0$ and

$$(3.22) \quad (t^k, p^k) \in \mathcal{Y}_{\epsilon^k}, \quad \rho_{\epsilon^k, \gamma^k}(t^k, p^k, u^k) \leq 1 \quad \forall k,$$

$$(3.23) \quad \|t^k\|/(\epsilon^k + \gamma^k) \rightarrow \infty, \quad \|t^k\| \rightarrow \infty.$$

By passing into a subsequence if necessary we will assume that $(t^k, p^k)/\|(t^k, p^k)\|$ converges to some limit point, say (t^∞, p^∞) (so $(t^\infty, p^\infty) \neq 0$). We have from (3.22) (also using (3.1) and (3.2)) that

$$Bt^k = d, \quad t^k > 0 \quad \forall k,$$

and from Lemma 3.1 that

$$b^T p^k \geq v^* - 2m\gamma^k, \quad \|t^k + A^T p^k - c\| \leq \sqrt{2(\eta^*(\epsilon^k)^2 + 2m\epsilon^k\gamma^k)} \quad \forall k.$$

Upon dividing both sides of the above four relations by $\|(t^k, p^k)\|$ and letting $k \rightarrow \infty$, we obtain from (3.23) and $(t^k, p^k)/\|(t^k, p^k)\| \rightarrow (t^\infty, p^\infty)$ that

$$Bt^\infty = 0, \quad t^\infty \geq 0, \quad b^T p^\infty \geq 0, \quad \|t^\infty + A^T p^\infty\| = 0.$$

Then, any optimal solution to (\mathcal{D}) would remain optimal when moved along the direction (t^∞, p^∞) , contradicting the boundedness of the optimal solution set of (\mathcal{D}) (via Assumption A(a)).

(b) Fix any $\epsilon > 0$, $\gamma > 0$, $\lambda \geq \epsilon\gamma/2$, and any $(t, p) \in \mathcal{Y}_\epsilon$ satisfying $\rho_{\epsilon, \gamma}(t, p, u) \leq 1$ for some u . Let $T = \text{Diag}(t)$ and let D, E, F be given by, respectively, (3.6), (3.7), and (3.8). Then, $F^{-1} = A(I - D^{1/2}ED^{1/2})A^T$, $\|E\| \leq 1$ and $D = (I + \lambda T^{-2})^{-1}$. From the definition of $\theta_\lambda(t)$ (see (3.5)) we then obtain

$$\begin{aligned} \theta_\lambda(t) &= (\|E + ED^{1/2}A^T FAD^{1/2}E\| + \|ED^{1/2}A^T F\|)^2 \\ &\leq (\|E\| + \|E\|^2\|D^{1/2}\|^2\|A\|^2\|F\| + \|E\|\|D^{1/2}\|\|A\|\|F\|)^2 \\ (3.24) \quad &< (1 + \|A\|^2\|F\| + \|A\|\|F\|)^2, \end{aligned}$$

where the strict inequality follows from the facts $\|D\| < 1$, $\|E\| \leq 1$. Now we bound $\|F\|$. We have

$$\begin{aligned} z^T(F^{-1})z &= z^T A(I - D^{1/2}ED^{1/2})A^T z \\ &\geq z^T A(I - D)A^T z \\ &= \sum_{j=1}^m (A_j^T z)^2 / ((t_j)^2 / \lambda + 1) \\ (3.25) \quad &\geq \lambda \sum_{j=1}^m (A_j^T z)^2 / (t_j)^2 \\ &\geq \lambda \sum_{j=1}^m (A_j^T z)^2 / \|t\|^2 \\ &\geq \lambda \sigma \|z\|^2 / \|t\|^2 \quad \forall z, \end{aligned}$$

where the first inequality follows from $\|E\| \leq 1$ and $\sigma > 0$ denotes the smallest eigenvalue of AA^T . ($\sigma > 0$ because A has full row rank.) Hence, $\lambda \geq \epsilon\gamma/2$ yields

$$(3.26) \quad \|F\| \leq 2\|t\|^2 / (\sigma\epsilon\gamma).$$

For ϵ and γ near zero, we give below a different bound on $\|F\|$. By the primal nondegeneracy assumption, there exists a constant $\delta > 0$ depending on A, B, b, c , and d only such that if (x, y) is any feasible solution of (\mathcal{P}) with $c^T x + d^T y \leq v^* + \delta$,

then the columns $\{A_j \mid j \in \{1, \dots, m\} \text{ with } x_j + B_j^T y \geq \delta\}$ have rank n . Consider the case in which $\eta^* \epsilon + 2m\gamma < \delta$, where η^* is defined as in Lemma 3.1. Since $(t, p) \in \mathcal{Y}_\epsilon$ and $\rho_{\epsilon, \gamma}(t, p, u) \leq 1$, we have from Lemma 3.1 that the columns $\{A_j \mid j \in \{1, \dots, m\} \text{ with } (t_j + A_j^T p + B_j^T u - c_j)/\epsilon \geq \delta\}$ have rank n . Since $(t, p) \in \mathcal{Y}_\epsilon$ and $\rho_{\epsilon, \gamma}(t, p, u) \leq 1$, we have $T(c - t - A^T p - B^T u) \geq -2\epsilon\gamma e$ (cf. (3.1) and (3.2)) so that $(t_j + A_j^T p + B_j^T u - c_j)/\epsilon \geq \delta$ implies $t_j \leq 2\gamma/\delta$. Hence, we obtain from (3.25) that

$$\begin{aligned} z^T(F^{-1})z &\geq \lambda \sum_{j=1}^m (A_j^T z)^2 / (t_j)^2 \\ &\geq \lambda(\delta)^2 \sum_{t_j \leq 2\gamma/\delta} (A_j^T z)^2 / (2\gamma)^2 \\ &\geq \lambda(\delta)^2 \sigma' \|z\|^2 / (2\gamma)^2 \quad \forall z, \end{aligned}$$

where the last inequality follows from the fact that those A_j for which $t_j \leq 2\gamma/\delta$ have rank n , and σ' is some positive scalar depending on A only. Since $\lambda \geq \epsilon\gamma/2$, we then have

$$\|F\| \leq (2\gamma)^2 / (\lambda(\delta)^2 \sigma') \leq 8\gamma / (\epsilon(\delta)^2 \sigma') = 8 / (\omega(\delta)^2 \sigma'),$$

where for convenience we let $\omega = \epsilon/\gamma$. Now, consider the remaining case where $\eta^* \epsilon + 2m\gamma \geq \delta$. Then, $\gamma \geq \delta/(2m)$, so part (a) and (3.26) together yield

$$\begin{aligned} \|F\| &\leq \frac{2(M_1)^2}{\sigma\epsilon\gamma} (\epsilon + \gamma + 1)^2 \\ &= \frac{2(M_1)^2}{\sigma} \left(\sqrt{\omega} + \frac{1}{\sqrt{\omega}} + \frac{1}{\gamma\sqrt{\omega}} \right)^2 \\ &\leq \frac{2(M_1)^2}{\sigma} \left(\sqrt{\omega} + \frac{1}{\sqrt{\omega}} + \frac{2m}{\delta\sqrt{\omega}} \right)^2. \end{aligned}$$

Combining the above two cases and we conclude that

$$\|F\| \leq \begin{cases} K/\omega & \text{if } \eta^* \epsilon + 2m\gamma < \delta; \\ K(\sqrt{\omega} + 1/\sqrt{\omega})^2 & \text{otherwise,} \end{cases}$$

for some positive scalar K depending on A, B, b, c , and d only. Combining the above bound with (3.24) and we conclude that $\theta_\lambda(t) \leq M_2(\sqrt{\omega} + 1/\sqrt{\omega})^4$ for some scalar $M_2 \geq 1$ depending on A, B, b, c , and d only. \square

3.3. Main convergence result. By combining Lemmas 3.1 to 3.3, we obtain the following global convergence result for the QLPPF Algorithm.

THEOREM 3.4. *Suppose that (\mathcal{P}) is primal nondegenerate in the sense of Lemma 3.3(b) and let $\psi(\cdot)$ be given by (3.21). If in the QLPPF Algorithm (t^1, p^1) together with some u^1 satisfies*

$$(3.27) \quad t^1 > 0, \quad Bt^1 = d,$$

$$(3.28) \quad \theta_{\epsilon^1, \gamma^1}(t^1) \leq \psi(\epsilon^1/\gamma^1),$$

$$(3.29) \quad \rho_{\epsilon^1, \gamma^1}(t^1, p^1, u^1) \leq \beta,$$

for some scalar

$$(3.30) \quad 0 < \beta < \frac{1}{\psi(\epsilon^1/\gamma^1)},$$

and if we choose

$$(3.31) \quad \alpha^k = \max \left\{ \sqrt{\frac{\theta_{\epsilon^k, \gamma^k}(t^k)(\beta)^2 + \sqrt{m}}{\beta + \sqrt{m}}}, \frac{1}{1 + \beta\sqrt{\gamma^1/\epsilon^1}/\|b\|} \right\} \quad \forall k,$$

then $\epsilon^k \downarrow 0$, $\gamma^k \downarrow 0$ linearly, and $\{((t^k + A^T p^k - c)/\epsilon^k, u^k/\epsilon^k)\}$, $\{(t^k, p^k)\}$ approach the optimal solution set of, respectively, (\mathcal{P}) and (\mathcal{D}) .

Proof. First note from $\psi(\omega) > 1$ for all $\omega > 0$ (see (3.21) and $M_2 \geq 1$) and (3.30) that $\beta < 1$. Also note from (2.7) that

$$(3.32) \quad \epsilon^k/\gamma^k = \epsilon^1/\gamma^1 \quad \forall k.$$

We claim that

$$(3.33) \quad (t^k, p^k) \in \mathcal{Y}_{\epsilon^{k-1}}, \quad \rho_{\epsilon^{k-1}, \gamma^{k-1}}(t^k, p^k, u^k) \leq \beta, \quad \rho_{\epsilon^k, \gamma^k}(t^k, p^k, u^k) \leq \beta,$$

for all $k \geq 2$. It is easily seen by using (3.27)–(3.31) and (2.5)–(2.7) and Lemma 3.2 that (3.33) holds for $k = 2$. Suppose that (3.33) holds for all $k \leq h$, for some $h \geq 2$. Then, $(t^h, p^h) \in \mathcal{Y}_{\epsilon^{h-1}}$ and $\rho_{\epsilon^{h-1}, \gamma^{h-1}}(t^h, p^h, u^h) \leq \beta < 1$. Since we also have from (2.7) that $\epsilon^h = \alpha^{h-1}\epsilon^{h-1}$ and $\gamma^h = \alpha^{h-1}\gamma^{h-1}$ and from (3.31) that $(\alpha^{h-1})^2 \geq \sqrt{m}/(1 + \sqrt{m}) \geq 1/2$, we can apply Lemma 3.3(b) to conclude that

$$(3.34) \quad \theta_{\epsilon^h, \gamma^h}(t^h) \leq \psi(\epsilon^{h-1}/\gamma^{h-1}) = \psi(\epsilon^1/\gamma^1),$$

where the equality follows from (3.32). Then, by (3.30), $\beta < 1/\theta_{\epsilon^h, \gamma^h}(t^h)$. Since (3.33) holds for $k = h$, we also have $\rho_{\epsilon^h, \gamma^h}(t^h, p^h, u^h) \leq \beta$, so Lemma 3.2 together with (2.5)–(2.7) and (3.31), (3.32) yields

$$(t^{h+1}, p^{h+1}) \in \mathcal{Y}_{\epsilon^h}, \quad \rho_{\epsilon^h, \gamma^h}(t^{h+1}, p^{h+1}, u^{h+1}) \leq \beta, \quad \rho_{\epsilon^{h+1}, \gamma^{h+1}}(t^{h+1}, p^{h+1}, u^{h+1}) \leq \beta.$$

Hence, (3.33) holds for $k = h + 1$.

Since (3.33) holds for all $k \geq 2$, we see that (3.34) holds for all $h \geq 2$. Then, by (3.30), $\theta_{\epsilon^h, \gamma^h}(t^h)\beta$ is less than 1 and bounded away from 1 for all $h \geq 2$, so that, by (3.31), α^k is less than 1 and bounded away from 1 for all k . Hence $\epsilon^k \downarrow 0$, $\gamma^k \downarrow 0$ at the rate of a geometric progression. The remainder of the proof follows from (3.33) and Lemma 3.1. \square

Note that instead of $\theta_{\epsilon^k, \gamma^k}(t^k)$ we can use, for example, the upper bound $1/\psi(\epsilon^1/\gamma^1)$ in the formula (3.31), and linear convergence would be preserved. However, this bound is typically loose and difficult to compute. There is also the issue of finding ϵ^1 , γ^1 , β , (t^1, p^1) and u^1 satisfying (3.27)–(3.30), which we will address in §3.4.

3.4. Algorithm initialization. By Theorem 3.4, if the primal nondegeneracy assumption therein holds and if we can find ϵ , γ , (t, p) and u satisfying

$$(3.35) \quad t > 0, \quad Bt = d,$$

$$(3.36) \quad \theta_{\epsilon, \gamma}(t) \leq \psi(\epsilon/\gamma),$$

$$(3.37) \quad \rho_{\epsilon, \gamma}(t, p, u) < 1/\psi(\epsilon/\gamma),$$

then we can set $\beta = \rho_{\epsilon, \gamma}(t, p, u)$ (assuming $\rho_{\epsilon, \gamma}(t, p, u) \neq 0$) and start the QLPPF Algorithm with ϵ , γ , (t, p) , and we would obtain linear convergence. But how do we find such ϵ , γ , (t, p) , and u ?

One obvious way is to fix any $\epsilon > 0$, any $\gamma > 0$, and then solve the penalized problem $(\mathcal{D}_{\epsilon,\gamma})$. The solution (t, p) obtained satisfies

$$t > 0, \quad Bt = d, \quad \nabla f_{\epsilon,\gamma}(t, p) + \begin{pmatrix} B^T u \\ 0 \end{pmatrix} = 0$$

for some u (see (2.2)) so, by (2.3) and (3.1), (3.2), we have $\rho_{\epsilon,\gamma}(t, p, u) = 0$ and $(t, p) \in \mathcal{Y}_\epsilon$. Hence (3.35) and (3.37) hold and, by Lemma 3.3(b), $\theta_{\epsilon,\gamma}(t) \leq \psi(\epsilon/\gamma)$, so (3.36) also holds. (Of course $(\mathcal{D}_{\epsilon,\gamma})$ need not be solved exactly.) To solve the problem $(\mathcal{D}_{\epsilon,\gamma})$, we can use any method for convex differentiable minimization (e.g., gradient descent, coordinate descent), and we would typically want ϵ small and γ large so that $(\mathcal{D}_{\epsilon,\gamma})$ is well conditioned.

Suppose that there holds $Be = 0$ and $Ae = b$. (This holds, for example, when B is the zero matrix (which corresponds to the case when all equality constraints are penalized), and a change of variable $x' = (\bar{X})^{-1}x$, where \bar{x} is any interior feasible solution of (\mathcal{P}) (i.e., $A\bar{x} = b$, $\bar{x} > 0$) and $\bar{X} = \text{Diag}(\bar{x})$, has been made in (\mathcal{P}) .) Then we can find a usable ϵ , γ , (t, p) , u immediately: Fix any $\epsilon > 2\psi(1)(\|c\| + \|B^T(BB^T)^{-1}d\|)$ and let $\gamma = \epsilon$. Also let $w = B^T(BB^T)^{-1}d$ and

$$\begin{aligned} p &= (AA^T)^{-1}A(c - w), \\ t &= \epsilon e + w, \\ u &= -(BB^T)^{-1}d. \end{aligned}$$

Then $Bw = d$, $A(c - A^T p) = Aw$, $At = \epsilon b + Aw$ and $B^T u = \epsilon e - t$, so that

$$\begin{aligned} Bt &= Bw = d, \\ A(c - t - A^T p) &= -\epsilon b, \\ T(c - t - A^T p - B^T u) + \epsilon \gamma e &= (\epsilon I + W)(c - \epsilon e - A^T p) + (\epsilon)^2 e \\ &= (\epsilon I + W)(c - A^T p) - \epsilon w, \end{aligned}$$

where $T = \text{Diag}(t)$ and $W = \text{Diag}(w)$. Also from $\psi(1) > 1$ and our choice of ϵ we have $\epsilon > \|w\|$, so $t > 0$. Hence, by (3.2), we have $(t, p) \in \mathcal{Y}_\epsilon$ and, by (3.1) and $\epsilon = \gamma$, we have

$$\begin{aligned} (3.38) \quad \rho_{\epsilon,\gamma}(t, p, u) &= \|(\epsilon I + W)(c - A^T p) - \epsilon w\|/(\epsilon)^2 \\ &\leq \|c - w - A^T p\|/\epsilon + \|W\|\|c - A^T p\|/(\epsilon)^2 \\ &= \|(I - A^T(AA^T)^{-1}A)(c - w)\|/\epsilon \\ &\quad + \|W\|\|(I - A^T(AA^T)^{-1}A)c + A^T(AA^T)^{-1}Aw\|/(\epsilon)^2 \\ &\leq \|c - w\|/\epsilon + \|w\|(\|c\| + \|w\|)/(\epsilon)^2, \end{aligned}$$

where the last inequality follows from the triangle inequality and the nonexpansive property of projection matrices. From our choice of ϵ we see that $(\|c\| + \|w\|)/\epsilon < .5/\psi(1)$, so the right-hand side of (3.38) is bounded above by $.5/\psi(1) + (.5/\psi(1))^2 \leq 1/\psi(1) = 1/\psi(\epsilon/\gamma)$, where the inequality follows from $\psi(1) \geq 1$ and the equality follows from $\gamma = \epsilon$. Hence (3.35) and (3.37) hold. Also, since $(t, p) \in \mathcal{Y}_\epsilon$, then (3.37) together with Lemma 3.3(b) shows that (3.36) holds.

4. Numerical results. To study the performance of the QLPPF Algorithm in practice, we have implemented the algorithm to solve the special case of (\mathcal{D}) and (\mathcal{P})

in which B is the zero matrix, i.e., (\mathcal{P}) is of the form

$$(4.1) \quad \begin{aligned} & \text{minimize } c^T x \\ & \text{subject to } Ax = b, \quad x \geq 0. \end{aligned}$$

(This corresponds to penalizing all equality constraints in the corresponding dual problem.) Below we describe our implementation and present our very preliminary numerical experience.

1. Initialization. In our implementation, we set for all problems

$$\epsilon^1 = 10^{-7} \|c\|_1/m, \quad \gamma^1 = 10^4 \|c\|_1/m,$$

and set (somewhat arbitrarily)

$$p^1 = 0, \quad t^1 = \max\{e, c/2\},$$

where “max” is taken componentwise. (Note that since B is the zero matrix, t^1 can be set to any positive vector.) We have chosen t^1 so to minimize $\|T^1(c - t^1 - A^T p^1)\|$ subject to $t^1 \geq e$. Also, care must be exercised in choosing ϵ^1 and γ^1 : if their values are set too low, then the QLPPF Algorithm may fail to converge; if their values are set too high, then the QLPPF Algorithm may require many iterations to converge. (Notice that we set ϵ^1 and γ^1 directly proportional to the average cost $\|c\|_1/m$ so their values scale with c .)

2. Steplength selection. To ensure that the t^k 's remain inside the positive orthant, we employ a backtracking scheme similar to that used by Setiono: whenever $t^k + \Delta t^k$ is outside the positive orthant, we replace the formula for t^{k+1} in (2.6) by

$$(4.2) \quad t^{k+1} = t^k + .98\lambda^k \Delta t^k,$$

where

$$(4.3) \quad \lambda^k = \min_{\Delta t_j^k < 0} -\frac{t_j^k}{\Delta t_j^k}.$$

However, this raises a difficulty, namely, for λ^k much smaller than 1, the vector $(.98\lambda^k \Delta t^k, \Delta p^k)$ may be far from the Newton direction $(\Delta t^k, \Delta p^k)$ and, as a consequence, the iterates may fail to converge. To remedy this, we replace (analogous to (4.2)) the formula for p^{k+1} in (2.6) by

$$p^{k+1} = p^k + .98\lambda^k \Delta p^k,$$

whenever nonconvergence is detected. (The parameter value .98 is chosen somewhat arbitrarily, but it works well in our tests.)

The proper choice of the α^k 's is very important for the QLPPF Algorithm: if the α^k 's are too near 1 (so the penalty parameters decrease slowly), then the algorithm would converge slowly; if the α^k 's are too near 0 (so the penalty parameters decrease rapidly), then the algorithm may fail to converge. In our implementation we adjust the α^k 's dynamically according to the following rule:

$$\alpha^k = \begin{cases} \max\{.3, .95\alpha^{k-1}\} & \text{if } \lambda^k = 1; \\ .6 & \text{if } \lambda^k \leq .2; \\ \alpha^{k-1} & \text{otherwise} \end{cases} \quad \forall k \geq 2,$$

with α^1 set to .5. The rationale for this adjustment rule is that, if $\lambda^k = 1$, then the current iterate is closely following the solution trajectory (so we can decrease the penalty parameters at a faster rate and still retain convergence) and, if $\lambda^k \leq .2$, then the current iterate is unable to follow the solution trajectory (so we must decrease the penalty parameters at a slower rate).

3. Termination. To avoid numerical problems, we stop decreasing the penalty parameters ϵ and γ when they reach some prespecified tolerances ϵ_{\min} and γ_{\min} , respectively. In our tests we set

$$\epsilon_{\min} = 10^{-12} \|c\|_1/m, \quad \gamma_{\min} = 10^{-9} \|c\|_1/m.$$

We terminate the QLPPF Algorithm when the relative duality gap and the violation of primal and dual feasibility are small. More specifically, we terminate whenever the current iterate, denoted by (t, p) , satisfies

$$(4.4) \quad \frac{\|X(A^T p - c)\|_1}{|c^T x|} \leq 10^{-7},$$

$$(4.5) \quad \max\{\|Ax - b\|_\infty, \|[-x]_+\|_\infty\} \leq 10^{-7},$$

$$(4.6) \quad \|[A^T p - c]_+\|_\infty \leq 10^{-7},$$

where $x = (t + A^T p - c)/\epsilon$ (see Lemma 3.1), $X = \text{Diag}(x)$, and $[\cdot]_+$ denotes the orthogonal projection onto the nonnegative orthant. Only for three of our test problems could the above termination criterion not be met (owing to violation of (4.4) and (4.6)) in which case the algorithm is terminated whenever primal feasibility (4.5) is met and $|c^T x - v^*|/|v^*|$ is less than $5 \cdot 10^{-7}$, where v^* denotes the optimal cost of (4.1).

4. Solving for the direction. The most expensive computation at each iteration of the QLPPF Algorithm lies in solving the system of linear equations (2.5). This can be seen to entail solving a single linear system of the form

$$(4.7) \quad AQA^T w = z,$$

for w , where z is some n -vector and Q is some $m \times m$ diagonal matrix whose j th diagonal entry is

$$(4.8) \quad \frac{\epsilon\gamma}{\epsilon\gamma + (t_j)^2},$$

with $\epsilon > 0$, $\gamma > 0$, and t some positive m -vector. (Linear system of the form (4.7) also arise in interior point algorithms, but (4.7) has the nice property that the condition number of Q can be controlled by adjusting the penalty parameters ϵ and γ .) In our implementation, (4.7) is solved using YSMP, a sparse matrix package for symmetric positive semidefinite systems developed at Yale University (see [EGSS79], [EGSS82]) and a precursor to the commercial package SMPAK (Scientific Computing Associates, 1985). YSMP comprises a set of Fortran routines implementing the Cholesky decomposition scheme and, as a preprocessor, the minimum-degree ordering algorithm (see, e.g., [GeL81]). In our implementation, the minimum-degree ordering routine ORDER is called first to obtain a permutation of the rows and columns of the matrix AA^T so that fill-in is reduced during factorization. Then, AA^T is symbolically factored using the routine SSF. (SSF is called only once since the nonzero pattern of AQA^T does not change with Q .) At each iteration, the matrix AQA^T is numerically factored by

the routine SNF (taking advantage of information generated by SSF concerning the location of the nonzeros in the factorization), and the two triangular systems thus generated are solved by the routine SNS to obtain a solution of (4.7). (We also experimented with the public domain version of the sparse matrix package SPARSPAK [GeL81], presently available from Netlib. We found SPARSPAK to be comparable to YSMP in solution time but somewhat inferior in solution accuracy.)

5. Data structure. The data structure used in our implementation is similar to that described in [AKRV89] and [MoM87]. Each matrix is stored in sparse format by row. To compute the nonzero entries of the matrix AQA^T efficiently for any Q , we also store the nonzero entries of the outer products $A_j(A_j)^T$, where A_j denotes the j th column of A . AQA^T is then computed using the formula

$$AQA^T = \sum_{j=1}^m q_j A_j(A_j)^T,$$

where q_j denotes the j th diagonal entry of Q and the product of q_j is taken with each nonzero entry of $A_j(A_j)^T$.

6. Test problems. Our test problems comprise the first 25 of the Netlib linear programming problems (see [Gay86]) used in the test of Monma and Morton [MoM87].² These problems range in size from 27 rows and 51 columns up to 1042 rows and 2869 columns and, for some of them, slack columns must be added and null rows must be removed to transform them into the form (4.1). (We also wrote a routine to convert these problems from their original MPS format to that used by our implementation.) The statistics for the test problems (after problem transformation) are summarized in Table 1.

7. Computing environment. Our implementation was written in Fortran and was compiled and ran on a Decstation 5000 under the Ultrix 4.2 operating system (similar to Berkeley 4.2 with some 4.3 enhancements). The default optimization setting for the compiler was used.

Table 2 summarizes the computational results obtained with our implementation of the QLPPF algorithm. Columns 2 and 3 show, respectively, the total number of iterations and the CPU time. Columns 4 and 6 show, respectively, the cost and the accuracy of the final primal solution (the latter is measured by the left-hand quantity in (4.5)). Analogously, columns 5 and 7 show, respectively, the cost and the accuracy of the final dual solution (the latter is measured by the left-hand quantity in (4.6)). For most of the problems, the primal cost agrees with the optimal cost in the first 7 digits and the accuracy of the primal solution is between 10^{-7} and 10^{-14} . Thus the quality of the computed solutions compares favorably with that of solutions generated by interior point algorithms. The number of iterations varies between 29 and 48 (except for Scagr25 which required 62 iterations) and the CPU time varies between 0.3 and 22 seconds, depending on the problem size and the sparsity of the constraint matrix. For most of the problems, over half of the CPU time is devoted to solving the linear system (4.7) at every iteration. (We also performed tests on a μ VAX-2000 Workstation under the operating system VMS 4.1. The resulting number of iterations is roughly the same; the accuracy of the final solutions improves slightly; and the CPU times are from 10 to 15 times that on the Decstation 5000.)

The number of iterations for the QLPPF Algorithm is comparable to that for the projected Newton barrier method of Gill et al. [GMSTW86], but is typically more

² Possibly due to conversion error, our version of Scorpion has an optimal cost very different from that reported in [Gay86] and hence the problem is excluded.

TABLE 1
Test problem characteristics

Problem Name	Number of Rows	Number of Cols	Constraint Nonzeros ¹	Hessian Nonzeros ²	Optimal Cost ³
Afiro	27	51	102	90	-4.6475314E+2
Adlittle	56	138	424	384	2.2549496E+5
Scagr7	129	185	465	629	-2.3313892E+6
Sc205	205	317	665	656	-5.2202061E+1
Share2B	96	162	777	871	-4.1573224E+2
Share1B	117	253	1179	1001	-7.6589319E+4
Scagr25	471	671	1725	2393	-1.4753433E+4
ScTap1	300	660	1872	1686	1.4122500E+3
BrandY	193	303	2202	2734	1.5185099E+3
Scsd1	77	760	2388	1133	8.6666666E+0
Israel	174	316	2443	11227	-8.9664482E+5
BandM	305	472	2494	3724	-1.5862801E-2
Scfxm1	330	600	2732	3233	1.8416759E+4
E226	223	472	2768	2823	-1.8751929E+1
Scrs8	490	1275	3288	2198	9.0429695E+2
Beaconfd	173	295	3408	2842	3.3592486E+4
Scsd6	147	1350	4316	2099	5.0500000E+1
Ship04s	360	1506	4400	3272	1.7987147E+6
Scfxm2	660	1200	5469	6486	3.6660261E+4
Ship04l	360	2166	6380	4588	1.7933245E+6
Ship08s	712	2467	7194	5440	1.9200982E+6
ScTap2	1090	2500	7334	6595	1.7248071E+3
Scfxm3	990	1800	8206	9739	5.4901254E+4
Ship12s	1042	2869	8284	6387	1.4892361E+6
Scsd8	397	2750	8584	4280	9.0499999E+2

¹ The number of nonzero entries in A .

² The number of nonzero entries in AA^T .

³ Cited from [Gay86].

than that for the affine-scaling algorithm or for Setiono's algorithm. Specifically, by comparing column 3 of [MoM87, Table 5] (also see [BDDW89], [MSSPB89]) with column 2 of Table 2, we see that the number of iterations for the QLPPF Algorithm can be up to $\frac{3}{2}$ times that for the affine-scaling algorithm. Similarly, the number of iterations for the QLPPF Algorithm can be up to $\frac{5}{3}$ times that for Setiono's algorithm (compare column 3 of [Set89, Table 3] with column 2 of Table 2). On the other hand, there are some problems on which the number of iterations is less for the QLPPF Algorithm than for the other algorithms.

In conclusion, our computational results indicate that, for linear programming, a mixed interior point-exterior point penalty method, as exemplified by the QLPPF Algorithm, can perform near the level of interior point algorithms. On the other hand, we caution that these results are very preliminary and thus should be viewed only as encouraging. In particular, the performance of the QLPPF Algorithm can be very sensitive to the choice of the initial parameters and the initial iterate. Of course, this also gives us hope that the performance of the QLPPF Algorithm can be improved with further fine tuning.

Finally, we remark it is typically beneficial to operate the QLPPF Algorithm with a large ratio of γ^k/ϵ^k . An intuitive explanation for this is that, if γ^k/ϵ^k is small, then γ^k is not a sufficiently large penalty (relative to ϵ^k) to maintain $t^k + \Delta t^k$ within the positive orthant. This results in small stepsizes λ^k (cf. (4.3)) and hence slow convergence.

TABLE 2
Computational results for the QLPPF Algorithm.

Problem Name	Iters.	CPU (sec.) ¹	Primal Cost ²	Dual Cost ²	Primal Feas. ³	Dual Feas. ³
Afiro	29	0.30	-4.6475313E+2	-4.6475316E+2	2E-08	0
Adlittle	37	0.64	2.2549499E+5	2.2549497E+5	3E-11	8E-09
Scagr7	43	0.98	-2.3313889E+6	-2.3313912E+6	3E-10	0
Sc205	36	1.07	-5.2202055E+1	-5.2202058E+1	1E-10	6E-12
Share2B	33	0.99	-4.1573226E+2	-4.1573228E+2	1E-10	0
Share1B	38	1.48	-7.6589327E+4	-7.6585942E+4	8E-08	2E-06
Scagr25	62	6.39	-1.4753429E+7	-1.4753437E+7	3E-09	9E-07
ScTap1	40	2.67	1.4122500E+3	1.4122499E+3	3E-12	0
BrandY	35	3.62	1.5185099E+3	1.5185098E+3	6E-08	1E-11
Scsd1	37	2.65	8.6666685E+0	8.6666668E+0	2E-15	0
Israel	44	21.10	-8.9664482E+5	-8.9659623E+5	4E-08	5E-05
BandM	37	4.31	-1.5862801E+2	-1.5862802E+2	4E-10	9E-10
Scfxm1	36	4.78	1.8416759E+4	1.8416758E+4	2E-08	3E-09
E226	43	4.71	-1.8751926E+1	-1.8751927E+1	7E-12	8E-11
Scrs8	48	7.43	9.0429699E+2	9.0429694E+2	6E-12	1E-08
Beaconfd	29	3.78	3.3592487E+4	3.3592485E+4	1E-09	3E-09
Scsd6	39	3.87	5.0500003E+1	5.0500000E+1	3E-13	0
Ship04s	35	4.43	1.7987148E+6	1.7987147E+6	8E-09	9E-09
Scfxm2	36	10.15	3.6660263E+4	3.6660260E+4	6E-08	3E-09
Ship04l	35	6.32	1.7933246E+6	1.7933245E+6	1E-10	0
Ship08s	40	9.68	1.9200982E+6	1.9200981E+6	2E-12	0
ScTap2	45	22.52	1.7248071E+3	1.7248071E+3	6E-13	3E-10
Scfxm3	36	17.93	5.4901256E+4	5.4901252E+4	2E-08	3E-09
Ship12s	38	12.44	1.4892362E+6	1.4892361E+6	1E-11	0
Scsd8	38	8.80	9.0500002E+2	9.0499999E+2	2E-13	0

¹ Obtained using the intrinsic function SECNDS on the Decstation 5000; does not include time to read the problem.

² Shown first 8 digits only.

³ Shown first digit only.

5. Some extensions. We have thus far assumed that the parameters ϵ and γ are decreased at the same rate in the QLPPF Algorithm. Alternatively we can decrease them at different rates. For example, Setiono's algorithm employs the strategy whereby ϵ is first decreased with γ held fixed and, once ϵ reaches a prescribed tolerance, then γ is decreased with ϵ held fixed. (In [Set89], the product $\epsilon\gamma$ is what is referred to as γ .) We can also use different penalty parameters for different coordinates.

Our convergence results very possibly also extend to linear complementarity problems with positive semi-definite matrices—in the same manner that the results in [Tse89] can be extended to these problems (see [Tse92]). This is a topic for further study.

Acknowledgment. I thank Dr. R. Setiono for providing me with a copy of the YSMP code and for helpful discussions regarding implementation. I also thank Professor D. P. Bertsekas for his support of this project and the referees for their helpful comments.

REFERENCES

- [AKRV89] I. ADLER, N. KARMARKAR, M. G. C. RESENDE, AND G. VEIGA, *Data structures and programming techniques for the implementation of Karmarkar's algorithm*, ORSA J. Comput., 1 (1989), pp. 84–106.

- [Ber82] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, NY, 1982.
- [BDDW89] P. T. BOGGS, P. D. DOMICH, J. R. DONALDSON, AND C. WITZGALL, *Algorithmic enhancements to the method of centers for linear programming problems*, ORSA J. Comput., 1 (1989), pp. 159–171.
- [EGSS79] S. C. EISENSTAT, M. C. GURSKY, M. H. SCHULTZ, AND A. H. SHERMAN, *Yale Sparse Matrix Package I. The Symmetric Codes*, Research Report #112, Department of Computer Science, Yale University, New Haven, CT (1979).
- [EGSS82] ———, *Yale Sparse Matrix Package I. The symmetric codes*, Internat. J. Numer. Methods Engrg., 18 (1982), pp. 1145–1151.
- [FiM68] A. V. FIACCO, AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, NY, 1968.
- [Fri57] K. R. FRISCH, *The Logarithmic Potential Method of Convex Programming*, Memorandum, University Institute of Economics, Oslo, Norway, May, 1955.
- [GaZ81] C. B. GARCIA, AND W. I. ZANGWILL, *Pathways to Solutions, Fixed Points, and Equilibria*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
- [Gay86] D. M. GAY, *Electronic Mail Distribution of Linear Programming Test Problems*, Manuscript 86-0, Department of Numerical Analysis, AT&T Bell Laboratories, Murray Hill, NJ, August, 1986.
- [GeL81] J. A. GEORGE AND J. W.-H. LIU, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
- [GMSTW86] P. E. GILL, W. MURRAY, M. A. SAUNDERS, J. A. TOMLIN, AND M. H. WRIGHT, *On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method*, Math. Prog., 36 (1986), pp. 183–209.
- [Gon89] C. C. GONZAGA, *An algorithm for solving linear programming problems in $O(n^3 L)$ operations*, in Progress in Mathematical Programming: Interior-Point and Related Methods, N. Megiddo, Ed., Springer-Verlag, New York, 1989, pp. 1–28.
- [HaB70] P. C. HAARHOFF AND J. D. BUYS, *A new method for the optimization of a nonlinear function subject to nonlinear constraints*, Computer J., 13 (1970), pp. 178–184.
- [Hes69] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
- [Hua67] P. HUARD, *Resolution of mathematical programming with nonlinear constraints by the method of centers*, in Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1967, pp. 207–219.
- [JiO78] K. JITTORNTRUM, AND M. R. OSBORN, *Trajectory analysis and extrapolation in barrier function methods*, J. Austral. Math. Soc. Series B, 20 (1978), pp. 352–369.
- [Kar84] N. KARMARKAR, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.
- [KMY89] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Prog., 44 (1989), pp. 1–26.
- [KMM91] M. KOJIMA, N. MEGIDDO, AND S. MIZUNO, *A Primal-Dual Exterior Point Algorithm for Linear Programming*, Technical Report, Mathematics and Computer Science Department, IBM Almaden Research Center, Almaden, CA, November, 1991.
- [Lus89] H. LUSS, ED., *Special Issue on Optimization*, AT&T Technical Journal, 68 (1989).
- [Lus90] I. J. LUSTIG, *Feasibility issues in a primal-dual interior-point method for linear programming*, Math. Prog., 49 (1990), pp. 145–162.
- [Man84] O. L. MANGASARIAN, *Normal solutions of linear programs*, Math. Prog. Study, 22 (1984), pp. 206–216.
- [MSSPB89] R. E. MARSTEN, M. J. SALTZMAN, D. F. SHANNO, G. S. PIERCE, AND J. F. BALLINTJN, *Implementation of a dual affine interior point algorithm for linear programming*, ORSA J. Comput., 1 (1989), pp. 287–297.
- [MMS89] K. A. MCSHANE, C. L. MONMA, AND D. SHANNO, *An implementation of a primal-dual interior point method for linear programming*, ORSA J. Comput., 1 (1989), pp. 70–83.
- [MeS90] S. MEHROTRA, AND J. SUN, *An algorithm for convex quadratic programming that requires $O(n^{3.5} L)$ arithmetic operations*, Math. Oper. Res., 15 (1990), pp. 342–363.
- [Meg89] N. MEGIDDO, ED., *Progress in Mathematical Programming: Interior-Point and Related Methods*, Springer-Verlag, New York, NY, 1989.
- [MoM87] C. L. MONMA AND A. J. MORTON, *Computational experience with a dual affine variant of Karmarkar's method for linear programming*, Oper. Res. Letters, 6 (1987), pp. 261–267.
- [MoA87] R. C. MONTEIRO AND I. ADLER, *Interior path following primal-dual algorithms - part I: Linear programming*, Math. Prog., 44 (1989), pp. 27–41.

- [MuR90] J. M. MULVEY, AND A. RUSZCZYŃSKI, *A Diagonal Quadratic Approximation Method for Large Scale Linear Programs*, Technical Report, Department of Civil Engineering and Operations Research, Princeton University, NJ, September, 1990.
- [PaS82] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, NJ, 1982.
- [Pow69] M. J. D. POWELL, *A Method for Nonlinear Constraints in Minimization Problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
- [Ren88] J. RENEGAR, *A polynomial-time algorithm, based on Newton's method, for linear programming*, Math. Prog., 40 (1988), pp. 59–93.
- [Roc70] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [Roc76] ———, *Augmented Lagrangians and applications for the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
- [Set89] R. SETIONO, *An Interior Dual Proximal Point Algorithm for Linear Programs*, Technical report, Computer Sciences Department, University of Wisconsin, Madison, WI, September, 1989.
- [Set90] ———, *Least Norm and Proximal Point Interior Algorithms for Linear Programming*, Ph.D. thesis, Computer Sciences Department, University of Wisconsin, Madison, WI, August, 1990.
- [Set91] ———, *Interior Dual Least 2-Norm Algorithm for Linear Programs*, Technical Report, Computer Sciences Department, University of Wisconsin, Madison, WI (July 1991); SIAM J. Control Optim., 31 (1993), pp. 875–899.
- [Tod89] M. J. TODD, *Recent Developments and New Directions in Linear Programming*, in Mathematical Programming: Recent Developments and Applications, M. Iri and K. Tanabe, eds., KTK Scientific Publishers, Tokyo, 1989, pp. 109–154.
- [Tse89] P. TSENG, *A simple complexity proof for a polynomial-time linear programming algorithm*, Oper. Res. Letters, 8 (1989), pp. 155–159.
- [Tse92] ———, *Complexity analysis of a linear complementarity algorithm based on a Lyapunov function*, Math. Prog., 53 (1992), pp. 297–306.
- [Ye87] Y. YE, *Further Development on the Interior Algorithm for Convex Quadratic Programming*, Technical Report, Integrated Systems Inc., Santa Clara, CA, November, 1987.
- [WBD88] C. WITZGALL, P. T. BOGGS, AND P. D. DOMICH, *On the Convergence Behavior of Trajectories for Linear Programming*, Technical Report, National Bureau of Standards, Boulder, CO, August, 1988.

DISCRETE-TIME TRANSITIVITY AND ACCESSIBILITY: ANALYTIC SYSTEMS*

FRANCESCA ALBERTINI[†] AND EDUARDO D. SONTAG[‡]

Abstract. A basic open question for discrete-time nonlinear systems is that of determining when, in analogy with the classical continuous-time “positive form of Chow’s Lemma,” accessibility follows from transitivity of a natural group action.

This paper studies the problem and establishes the desired implication for analytic systems in several cases: (i) compact state space, (ii) under a Poisson stability condition, and (iii) in a generic sense. In addition, the paper studies accessibility properties of the “controllability sets” recently introduced in the context of dynamical systems studies. Finally, various examples and counterexamples are provided relating the various Lie algebras introduced in past work.

Key words. discrete-time, nonlinear systems, transitivity, accessibility

AMS subject classifications. primary 93B03, 93B05; secondary 93C55, 93B29

1. Introduction. This paper continues the study, initiated in [7], of systems of the type

$$(1) \quad x(t+1) = f(x(t), u(t)), \quad t = 0, 1, 2, \dots,$$

where x and u take values in manifolds. The smooth mapping f is assumed to be invertible on x for each fixed u , a restriction that models systems that arise when dealing with continuous-time plants under digital control. See [7] for further motivation for the study of such systems, and [12] for general definitions of systems.

Given the system (1), we may introduce the *reachable* or *forward-accessible set* from a state x^0 , which we will denote by $R(x^0)$. This is the set of states to which we may steer x^0 using arbitrary controls. Clearly, reachable sets are one of the central concepts in control theory.

A mathematically far easier object to deal with is the *orbit* or *forward-backward accessible set* from x^0 , which we will denote by $O(x^0)$. This is defined as the set consisting of all states to which x^0 can be steered using both motions of the system and negative time motions: a state z is in the orbit of x^0 if there exists a sequence of states

$$x_0 = x^0, x_1, \dots, x_k = z$$

such that, for each $i = 1, \dots, k$, either x_i is reachable from x_{i-1} or x_{i-1} is reachable from x_i .

Of course, $R(x^0)$ is always included in $O(x^0)$, but these two sets are in general different. Observe that $O(x^0)$ is the orbit of x^0 under the group action induced by all the diffeomorphisms $f(\cdot, u)$, while the main interest in control theory—since negative time motions are in general not physically realizable—is in $R(x^0)$, the orbit under the corresponding semigroup. One reason that orbits are easier to study is that they have

* Received by the editors April 22, 1991; accepted for publication (in revised form) April 9, 1992. This research was supported in part by U.S. Air Force Grant AFOSR-91-0346.

[†] SYCON–Rutgers Center for Systems and Control, Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903 (albertin@hilbert.rutgers.edu). Present address, Università di Padova, Dipartimento di Matematica, via Belzoni 7, 35100 Padova, Italy.

[‡] SYCON–Rutgers Center for Systems and Control, Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903 (sontag@hilbert.rutgers.edu).

a natural structure of submanifold of the state space; this induces a decomposition of the state space into invariant submanifolds that integrate a natural distribution of vector fields (see, for instance, [13] and [11]).

One of the central facts in continuous-time controllability is the following property, valid for analytic systems and arbitrary states x^0 :

$$(C) \quad R(x^0) \text{ has nonempty interior in } O(x^0).$$

This property follows directly from the orbit theorem, but it can also be established for general smooth systems, under appropriate Lie-algebraic assumptions; it is often known as the “positive form of Chow’s Lemma.” Thus, for continuous-time, the state space can be partitioned into invariant submanifolds, and inside each submanifold we can reach an open set from each state. In particular, the interior of the reachable set from x^0 is nonempty—we then say that there is *forward accessibility from x^0* —if and only if the orbit is open—i.e., there is *transitivity from x^0* .

In contrast, property (C) may fail in discrete-time, even for systems obtained through the time-sampling of one-dimensional analytic continuous-time systems; see the examples in [7]. There are two known cases where (C) does hold:

(a) When x^0 is an *equilibrium point* (and the system is analytic and the control-value set is connected); this is one of the main results in [7].

(b) If the map f is *rational* on states and controls; see [9].

Both of these properties are quite restrictive; equilibria are in general few, and the rationality assumption is too strong in discrete-time (note that even when sampling very simple—for instance, polynomial—continuous-time systems we do not in general obtain rational equations.)

In this paper we extend the validity of property (C). For analytic systems, we prove that property (C) does hold if the orbit from x^0 is compact (see Remark 4.1), or under certain stability hypotheses related to Hamiltonian dynamics. Another result shows that if there is only one orbit (the system is transitive), then forward accessibility holds from an *open dense* set of states, assuming the state space to have at most finitely many connected components. Building upon the results in this paper, [2] provides further conditions under which property (C) holds.

Low-dimensional cases are of interest because certain special implications hold in those cases, and as sources of examples and counterexamples. For instance, we show that in dimension one transitivity from a given state x^0 implies either forward accessibility from x^0 or backward accessibility (controllability from some open set to x^0), but that this result fails in dimension two.

Recently, Colonius and Kliemann introduced the notion of *controllability subsets* of the state space of continuous-time systems. These are essentially sets where “almost reachability” holds. Controllability sets have proved to be an extremely useful concept; in particular, in [4] these authors established an interesting relationship between such sets and chaotic behavior in subsets of an associated dynamical system. The extension to discrete-time of the results of Colonius and Kliemann depends critically on the better understanding of the forward accessibility properties of controllability sets, so we devote the last part of this paper to that goal. The reader is referred to the conference paper [1] for a detailed explanation of how the results in [4] can indeed be extended when applying the techniques developed here.

2. Basic definitions. In this paper we will deal with discrete-time nonlinear systems Σ of the type (1) where $x(t) \in X$ and $u(t) \in U$. We assume that the

state space X is a connected, second countable, Hausdorff, differentiable manifold of dimension n , except in §5.1, where we wish to study what happens if the connectedness assumption is dropped.

The control-value space U is always assumed to be a subset of \mathbb{R}^m that satisfies the assumptions

$$U \subseteq \text{clos int } U$$

and $0 \in U$. We always assume that U is a connected set, except in §§3.1 and 6 where this assumption can be dropped.

The system is of class C^k , with $k = \infty$ or ω , if the manifold X is of class C^k and the function

$$f : X \times U \rightarrow X$$

is of class C^k (i.e., there exists a C^k extension of f to an open neighbourhood of $X \times U$ in $X \times \mathbb{R}^m$). We call systems of class C^∞ *smooth systems* and those of class C^ω *analytic systems*.

The most restrictive technical assumption to be made is that the system is *invertible*; this means that for each $u \in U$ the map $f_u = f(\cdot, u) : X \rightarrow X$ is a global diffeomorphism of X . Invertibility allows the application of the techniques in [7]; the assumption is satisfied when dealing with systems obtained by sampling a continuous-time one. We will use f_u^{-1} to denote the inverse of the map f_u .

Unless otherwise stated, from now on we assume that a fixed smooth system Σ is given.

2.1. Some notation. If there exists an integer $k \geq 0$ and a k -tuple $(u_k, \dots, u_1) \in U^k$ such that $f_{u_k, \dots, u_1}(x) = z$, we will write

$$x \overset{\sim}{\underset{k}{\rightsquigarrow}} z.$$

As usual, f_{u_k, \dots, u_1} denotes $f_{u_k} \circ \dots \circ f_{u_1}$. For any fixed state x and any nonnegative integer k define:

$$\psi_{k,x}(\mathbf{u}) := f_{u_k, \dots, u_1}(x),$$

where $\mathbf{u} = (u_k, \dots, u_1) \in U^k$. For each \mathbf{u} , let $\rho_{k,x}(\mathbf{u})$ be the rank of $\partial/\partial \mathbf{u} \psi_{k,x}[\mathbf{u}]$, and denote

$$\bar{\rho}_{k,x} := \max_{\mathbf{u} \in U^k} \rho_{k,x}(\mathbf{u}).$$

For each x , let also

$$\bar{\rho}_x := \max_{k \geq 0} \bar{\rho}_{k,x};$$

roughly, this is the largest possible dimension of a manifold reachable from x . Observe that $k' \geq k$ implies

$$(2) \quad \bar{\rho}_{k',x} \geq \bar{\rho}_{k,x}$$

because if $\mathbf{u} \in U^k$ achieves $\rho_{k,x}(\mathbf{u}) = \bar{\rho}_{k,x}$ then also $\rho_{k',x}(\tilde{\mathbf{u}}) \geq \rho_{k,x}(\mathbf{u})$ for any $\tilde{\mathbf{u}} \in U^{k'}$ that extends \mathbf{u} . We define the following sets:

$$R^k(x) := \{z \mid x \overset{\sim}{\underset{k}{\rightsquigarrow}} z\}$$

is the set of states *reachable from x in (exactly) k steps*,

$$\tilde{R}^k(x) := \{\psi_{k,x}(\mathbf{u}) \mid \mathbf{u} \in U^k, \rho_{k,x}(\mathbf{u}) = \bar{\rho}_x\}$$

is the set of states that are *maximal-rank reachable from x in (exactly) k steps*,

$$\bar{R}^k(x) := \{\psi_{k,x}(\mathbf{u}) \mid \mathbf{u} \in U^k, \rho_{k,x}(\mathbf{u}) = n\}$$

is the set of states that are *nonsingularly reachable from x in k steps*. Observe that, clearly,

$$\bar{R}^k(x) \subseteq \tilde{R}^k(x) \subseteq R^k(x).$$

We let

$$R(x) := \bigcup_{k \geq 0} R^k(x)$$

and analogously for $\tilde{R}(x)$ and $\bar{R}(x)$. Recall that Σ is said to be *forward accessible from x* if and only if $\text{int } R(x) \neq \emptyset$.

We also define

$$\begin{aligned} O^0(x) &= x, \\ O^k(x) &= \{z \mid \exists z_1 \in O^{k-1} \text{ and } z_1 \rightsquigarrow_1 z \text{ or } z \rightsquigarrow_1 z_1\}, \end{aligned}$$

and

$$O(x) = \bigcup_{k \geq 0} O^k(x).$$

Thus $O(x)$ is the orbit from x ; Σ is said to be *transitive from x* if and only if $\text{int } O(x) \neq \emptyset$.

Note that, given any state x , there is a well-defined restriction of the system to the orbit $O(x)$. Hence all results can be, in principle, applied in each orbit. The only difficulty is that orbits are often *not* connected, while most results hold only under the blanket assumption that the state space must be connected. In §5.1 we make some further comments about this issue.

Certain Lie algebras of vector fields L , L^+ , Γ , Γ^+ were introduced in [7] (see also [5] and [6] for previous work) we repeat their definitions here for the convenience of the reader.

First we let X_u^+ and X_u^- be the following vector fields:

$$X_{u,i}^+(x) = \left. \frac{\partial}{\partial v_i} \right|_{v=0} f_u^{-1} \circ f_{u+v}(x),$$

$$X_{u,i}^-(x) = \left. \frac{\partial}{\partial v_i} \right|_{v=0} f_u \circ f_{u+v}^{-1}(x),$$

one for each $i = 1, \dots, m$ (for computational aspects associated to these vector fields see [3]). Given a vector field Y and a control value u , we can define another vector field from Y by applying a change of coordinates given by the diffeomorphism f_u ,

$$(\text{Ad}_u Y)(x) = (df_u(x))^{-1} Y(f_u(x)).$$

Here df_u stands for the differential of f_u with respect to x . In the same way, but now using the diffeomorphism f_u^{-1} , we also define Ad_u^{-1} . We let

$$(3) \quad \text{Ad}_{u_k \dots u_1}^{\epsilon_k \dots \epsilon_1} Y = \text{Ad}_{u_1}^{\epsilon_1} \dots \text{Ad}_{u_k}^{\epsilon_k} Y.$$

We will use the abbreviated notation $\text{Ad}_0^k Y$ for $\text{Ad}_{0 \dots 0} Y$ with $u = 0$ repeated k -times, if $k > 0$, and for $\text{Ad}_0^{-1} \dots \text{Ad}_0^{-1} Y$, if $k < 0$. Additionally, $\text{Ad}_0^0 Y = Y$. The Lie algebras Γ^+ and Γ are now defined as

$$\Gamma^+ = \{\text{Ad}_{u_k \dots u_1} X_{u_0, i}^+ \mid k \geq 0, 1 \leq i \leq m, u_0, \dots, u_k \in U\},$$

$$\Gamma = \{\text{Ad}_{u_k \dots u_1}^{\epsilon_k \dots \epsilon_1} X_{u_0, i}^\sigma \mid k \geq 0, 1 \leq i \leq m, u_0, \dots, u_k \in U, \epsilon_1, \dots, \epsilon_k = \pm 1, \sigma = \pm\}.$$

Finally the Lie algebras L^+ and L are as follows:

$$L^+ = \text{Lie} \{\text{Ad}_0^k X_{u, i}^+ \mid k \geq 0, 1 \leq i \leq m, u \in U\},$$

$$L = \text{Lie} \{\text{Ad}_0^k X_{u, i}^+ \mid k \in \mathbb{Z}, 1 \leq i \leq m, u \in U\}.$$

We look at the sets of states in which various rank conditions fail, or forward accessibility fails:

$$\begin{aligned} B^+ &:= \{x \mid \text{int } R(x) = \emptyset\} \\ B_L^+ &:= \{x \mid \dim L^+(x) < n\} \\ B_\Gamma^+ &:= \{x \mid \dim \Gamma^+(x) < n\}. \end{aligned}$$

Although well-defined always, the set B_L^+ will be of interest only when the system is analytic.

2.2. Review of main known facts. With this notation, many of the results obtained in [7] can be visualized by the following diagram, where an arrow “ $A \rightarrow B$ ” indicates inclusion $A \subseteq B$, and the inclusions involving B_L^+ are only valid in the analytic case:

$$\begin{array}{ccccc} \tilde{R}(B^+) & \longrightarrow & B_\Gamma^+ & \longrightarrow & B_L^+ \\ \downarrow & & \downarrow & \swarrow & \\ R(B^+) & \longrightarrow & B^+ & & \end{array}$$

Note. The inclusion

$$(4) \quad \tilde{R}(B^+) \subseteq B_\Gamma^+$$

rephrases the result obtained in Corollary 4.4 of [7]. The inclusion $B_\Gamma^+ \subseteq B^+$ expresses the result in Theorem 6 part (a) of [7]. The inclusion $B_L^+ \subseteq B^+$ represents the result in Theorem 6 part (b) of [7].

3. Some new general properties. In this section we prove a number of general facts that can be conveniently expressed in terms of the sets just defined.

Remark 3.1. If there exists any k_0 such that $\bar{R}^{k_0}(x)$ is nonempty, then for all $k \geq k_0$ we have $\bar{R}^k(x) = \tilde{R}^k(x)$. Indeed, the assumption implies that $\bar{\rho}_x = n$.

PROPOSITION 3.1. *For each $x \in X$, the following properties are equivalent:*

- (a) $\text{int } \bar{R}(x) \neq \emptyset$;
- (b) $\text{int } \tilde{R}(x) \neq \emptyset$;
- (c) $\text{int } R(x) \neq \emptyset$.

Proof. Since $\bar{R}(x) \subseteq \tilde{R}(x) \subseteq R(x)$, it is only necessary to show that (c) implies (a).

We will show the following two properties:

- 1. for each $k \geq 0$ if $\text{int } \bar{R}^k(x) = \emptyset$ then $\bar{R}^k(x) = \emptyset$;
- 2. if $\bar{R}^k(x) = \emptyset$ for all $k \geq 0$ then $\text{int } R(x) = \emptyset$.

Combining (1) and (2) we have that if $\text{int } \bar{R}(x) = \emptyset$ then all $\text{int } \bar{R}^k(x) = \emptyset$ too, so $\text{int } R(x) = \emptyset$, as desired.

We first prove (1). Suppose that $\bar{R}^k(x) \neq \emptyset$, so that there exists some sequence $\bar{\mathbf{u}}$ for which the rank $\rho_{k,x}(\cdot)$ is equal to n at $\bar{\mathbf{u}}$. Since we assume $U \subset \text{clos int } U$, there exists also some $\tilde{\mathbf{u}} \in \text{int } U^k$ so that $\rho_{k,x}(\mathbf{u}) = n$ for each \mathbf{u} in some neighbourhood of $\tilde{\mathbf{u}}$. By the implicit mapping theorem, $\tilde{z} = \psi_{k,x}(\tilde{\mathbf{u}})$ belongs to $\text{int } \bar{R}^k(x)$.

We now prove (2). If $\bar{R}^k(x) = \emptyset$ for all $k \geq 0$ then each $\mathbf{u} \in U^k$ is a singular point of the map $\psi_{k,x}$, for each k . Thus by Sard's theorem $\psi_{k,x}(U^k)$ has measure zero for all $k \geq 0$. It follows that also

$$R(x) = \bigcup_{k \geq 0} R^k(x) = \bigcup_{k \geq 0} \psi_{k,x}(U^k)$$

has measure zero, and hence $\text{int } R(x) = \emptyset$, as desired. \square

PROPOSITION 3.2. *If the system Σ is analytic then, for any $x \in X$:*

$$\text{clos } R^k(x) = \text{clos } \tilde{R}^k(x)$$

for all k sufficiently large.

Proof. Fix $x \in X$, and let k_0 be so that $\bar{\rho}_{k_0,x} = \bar{\rho}_x$. For all $k \geq k_0$, let

$$A_k(x) = \{\mathbf{u} \mid \rho_{k,x}(\mathbf{u}) = \bar{\rho}_x\}.$$

We claim that $A_k(x)$ is an open dense set of U^k . This is because $A_k(x) \neq \emptyset$ by (2) and the complement of $A_k(x)$ is a set defined by the vanishing of certain analytic functions (suitable determinants) of \mathbf{u} .

We claim that

$$R^k(x) \subseteq \text{clos } \tilde{R}^k(x),$$

which implies

$$(5) \quad \text{clos } R^k(x) \subseteq \text{clos } \tilde{R}^k(x) \text{ for each such } k.$$

This will establish the result, the other inclusion being obvious.

Indeed, pick $k \geq k_0$ and take any $z \in R^k(x)$. Then $z = \psi_{k,x}(\mathbf{u})$ for some $\mathbf{u} = (u_k, \dots, u_1)$. Since $A_k(x)$ is dense, we can find a sequence $\{\mathbf{u}_l\}$ such that

$$\mathbf{u}_l = (u_k^{(l)}, \dots, u_1^{(l)}) \rightarrow \mathbf{u} = (u_k, \dots, u_1) \text{ as } l \rightarrow \infty$$

and $\mathbf{u}_l \in A_k(x)$ for each l .

Let $z_l = \psi_{k,x}(\mathbf{u}_l) \in \tilde{R}^k(x)$. By continuity, $z_l \rightarrow z$, which proves (5). \square

Remark 3.2. Assume that the system Σ is analytic, and that there exists an $x_0 \in X$ and a $k_0 \geq 0$ for which $\bar{R}^{k_0}(x_0) \neq \emptyset$. Then the proof of the previous result together with Remark (3.1) imply that

$$\text{clos } R^k(x_0) = \text{clos } \bar{R}^k(x_0)$$

for all $k \geq k_0$.

Moreover, since $\partial/\partial \mathbf{u} \psi_{k,x}[\mathbf{u}]$ is analytic also with respect to the x -variable, this particular k_0 works also for an open dense set of states $x \in X$. Thus, under these assumptions, we have that

$$\text{clos } R^k(x) = \text{clos } \tilde{R}^k(x) = \text{clos } \bar{R}^k(x)$$

for all $k \geq k_0$ and for almost all $x \in X$.

3.1. Regular points. We call x a *regular point* if $\bar{\rho}_x$ is constant in a neighbourhood of x . The following fact will be useful later; it is of course a well-known general fact about smooth mappings.

LEMMA 3.3. *The regular points form an open dense subset of X .*

Proof. Let

$$\bar{\rho} = \max_{x \in X} \bar{\rho}_x.$$

We have $\bar{\rho} \in \{0, \dots, n\}$. We will prove our thesis by induction on $\bar{\rho}$.

If $\bar{\rho} = 0$, then each $x \in X$ is a regular point, thus the statement is true.

Let $\bar{\rho} > 0$. Define

$$\begin{aligned} X_1 &:= \{x \in X \mid x \text{ is a regular point and } \bar{\rho}_x = \bar{\rho}\}, \\ Y_1 &:= \text{int } \{X \setminus X_1\}. \end{aligned}$$

Then X_1 and Y_1 are open. Moreover $X_1 \cup Y_1$ is dense in X , since its complement is the boundary of X_1 which is a nowhere dense set. If we call

$$\bar{\rho}_1 = \max_{x \in Y_1} \bar{\rho}_x$$

we have $\bar{\rho}_1 < \bar{\rho}$.

Thus, applying the inductive assumption to $(\bar{\rho}_1, Y_1)$, we have that the set of regular points in Y_1 , denote it by Y_r , is dense in Y_1 . But since the set of regular points of X is given by $X_1 \cup Y_r$ and $X_1 \cup Y_1$ is dense in X , then $X_1 \cup Y_r$ is also dense in X . \square

Note that, in the particular case in which the system is analytic, then in the above proof the set X_1 is already dense, because the rank is less than $\bar{\rho}$ if and only if certain determinants, which are analytic functions of x , vanish and this can happen only in a nowhere dense set.

4. More results for analytic systems. In this section we always assume the system Σ to be *analytic*.

LEMMA 4.1. *Suppose that for a fixed $x \in X$ there exists a sequence of elements $\{x_{n_k}\}$ and some $y \in X$ so that $\dim L^+(y) = n$, such that*

1. $x_{n_k} \in R^{n_k}(x)$, with $n_k \rightarrow \infty$;
2. $x_{n_k} \rightarrow y$.

Then the system is forward accessible from x (i.e. $x \notin B^+$).

Proof. Since $x_{n_k} \rightarrow y$ and $\dim L^+(y) = n$ there is some integer $k_0 \geq 0$ such that $\dim L^+(x_{n_k}) = n$ for all $k \geq k_0$. But for k sufficiently large we know (by Proposition (3.2)) that $x_{n_k} \in \text{clos } \tilde{R}^{n_k}(x)$. Thus there exists some $z \in X$ such that $z \in \tilde{R}^{n_k}(x)$ and $\dim L^+(z) = n$. So we can conclude forward accessibility from x by (4). \square

Remarks 4.1.

1. The result is also true if the weaker assumption $\dim \Gamma^+(y) = n$ is made, but we will apply it in the above form.

2. If x and y are as in the previous lemma, and U is any open neighbourhood of y , then, in particular, we have that $R(x) \cap U$ is also open.

3. If for a fixed $x \in X$ there exists a sequence of elements $\{x_{n_k}\}$ such that $x_{n_k} \in R^{n_k}(x)$, with $n_k \rightarrow \infty$ and $x_{n_k} \rightarrow x$ then, by the previous lemma, we can conclude that forward accessibility from x is equivalent to $\dim L^+(x) = n$. We will see later that in dimension 1 this equivalence is always true, but it can fail in higher dimensions.

For each $x \in X$, we will denote by $y_{0,x}^k$ the image under $\psi_{k,x}(\cdot)$ of the zero control; i.e.,

$$y_{0,x}^k = \psi_{k,x}(\underbrace{0, \dots, 0}_{k\text{-times}}).$$

LEMMA 4.2. *Suppose that $x, y \in X$ are so that*

1. *the system is transitive from y , (or equivalently, $\dim L(y) = n$),*

2. *there exists a sequence $\{y_{0,x}^{n_k}\}$ with $n_k \rightarrow \infty$ such that $y_{0,x}^{n_k} \rightarrow y$.*

Then $\dim L^+(x) = n$.

Proof. Choose n vector fields v_1, \dots, v_n in L such that

$$\{v_1(y), \dots, v_n(y)\}$$

is a basis for $L(y)$.

As in the proof of Proposition 4.2 in [7], we can assume that the v_i 's involve Lie brackets of a finite numbers of vector fields of the form $Ad_0^{k_j} X_{u_j}^+$, with $k_j \in \mathbb{Z}$. Choose a positive integer k_0 so that $k_j + k_0 \geq 0$ for all such j .

Since the v_i 's are linearly independent at y , they are still linearly independent in some neighbourhood U_y of y . By assumption (2), there is some n_k so that $y_{0,x}^{n_k} \in U_y$ and $n_k \geq k_0$.

Applying the operator $Ad_0^{n_k}$ to the v_i 's, there result n linearly independent vectors in $L^+(x)$, as desired. \square

4.1. Poisson stability. Recall that if Y is a vector field on a manifold M , one says that $x \in M$ is a *positively Poisson stable point* for Y if and only if for each neighbourhood V of x and each $T \geq 0$ there exists some $t > T$ such that $e^{tY}(x) \in V$, where $e^{tY}(\cdot)$ represents the flow of Y .

Analogously, we can define positive Poisson stability in discrete time, as follows.

DEFINITION 4.1. Let $f : X \rightarrow X$ be a global diffeomorphism. The point $x \in X$ is *positively Poisson stable* if and only if for each neighbourhood V of x and each integer $N \geq 0$ there exists some integer $k > N$ such that $f^k(x) \in V$.

THEOREM 4.3. *Let $x \in X$ be a positively Poisson stable point for $f_0 = f(\cdot, 0)$. Then transitivity from x implies forward accessibility from x .*

Proof. Positive Poisson stability from x implies the existence of a sequence $\{y_{0,x}^{n_k}\}$, with $n_k \rightarrow \infty$, convergent to x . Thus the result follows immediately combining Lemmas (4.1), (4.2) (applied with $y = x$). \square

4.2. Compact state space. For each $k \geq 0$ we define the following sets:

$$C^k(x) := \{y \mid y \overset{\sim}{\rightsquigarrow}_k x\},$$

i.e., the set of states *controllable to* x in (exactly) k steps, and

$$C(x) = \bigcup_{k \geq 0} C^k(x).$$

A system is *backward accessible from* x if and only if $\text{int } C(x) \neq \emptyset$.

THEOREM 4.4. *Let Σ be a discrete time, analytic, invertible system, and assume that the state space X is compact.*

Then, Σ is transitive if and only if it is forward accessible.

Proof. By [7], Theorem 3, it will be enough to show that $\dim L^+(x) = n$ for all $x \in X$. Fix any $x \in X$, and consider the sequence

$$y_{0,x}^l = \psi_{l,x}(0, \dots, 0).$$

Then since X is compact (and second countable) there exists a subsequence $\{y_{0,x}^{l_k}\}$ which converges; let y be so that $y_{0,x}^{l_k} \rightarrow y$. Since Σ is transitive, $\dim L(y) = n$, so, by Lemma (4.2), $\dim L^+(x) = n$ as wanted. \square

Remark 4.1. Notice that, in the previous theorem, the blanket assumption of connectedness of the state space X is not needed. In particular, the result holds if the orbit from a state x is compact.

Remark 4.2. Clearly, using the same arguments as in Theorem 4.4, we also have that, if the state space is compact, then transitivity from all $x \in X$ is equivalent to backward accessibility from all $x \in X$. We will not use this fact, however.

Recall that for a space Z with a σ -algebra F and a finite measure μ , we say that a measurable transformation $T: Z \rightarrow Z$ is *measure-preserving* if for every $A \in F$ we have $\mu(T^{-1}A) = \mu(A)$.

The following controllability result is an analogue for discrete-time systems of the result in [8]. The proof is very similar, but it uses the facts just established.

PROPOSITION 4.5. *Assume that the state space X is a compact Riemannian analytic manifold, and that for all $u \in U$ the map f_u is a measure preserving transformation (for the natural measure in X). Then Σ is transitive if and only if Σ is controllable.*

Proof. We need only to prove that transitivity implies controllability.

For each u , since f_u is a measure preserving map, by the Poincaré recurrence theorem the set of positively Poisson stable points for f_u is known to be dense in X .

Let $x, y \in X$; we need $y \in R(x)$. By Theorem 4.4, we know that Σ is both forward and backward accessible from x and y . Choose $\bar{x} \in \text{int } R(x)$ and $\bar{y} \in \text{int } C(y)$; since Σ is transitive there exist k , (u_k, \dots, u_1) , and $(\epsilon_k, \dots, \epsilon_1)$, with each $u_i \in U$ and $\epsilon_i = 1$ or -1 , such that

$$f_{u_k}^{\epsilon_k} \circ \dots \circ f_{u_1}^{\epsilon_1}(\bar{x}) = \bar{y}.$$

Let $l = \text{number of } \epsilon_i = -1$. We will show by induction on l the following fact:

there exist $\tilde{x} \in \text{int } R(x)$ and $\tilde{y} \in \text{int } C(y)$ such that $\tilde{y} \in R(\tilde{x})$.

Clearly the previous statement implies our thesis.

If $l = 0$ then the statement holds with $\tilde{x} = \bar{x}$ and $\tilde{y} = \bar{y}$. So let $l > 0$ and let i be the first index such that $\epsilon_i = -1$. Define

$$x_i = f_{u_{i-1}} \circ \dots \circ f_{u_1}(\bar{x})$$

and

$$y_i = f_{u_i}^{-1}(x_i).$$

Since $\bar{y} \in \text{int } C(y)$, there exists a neighbourhood V of y_i such that

$$f_{u_k}^{\epsilon_k} \circ \dots \circ f_{u_{i+1}}^{\epsilon_{i+1}}(V) \subseteq C(y);$$

let $W = f_{u_i}(V)$. Since $\bar{x} \in \text{int } R(x)$ we can assume (taking V smaller if necessary) that $W \subseteq R(x)$.

Choose $z_i \in W$ positively Poisson stable for f_{u_i} ; then there exists some $n > 1$ such that $f_{u_i}^n(z_i) \in W$ and the following properties hold:

- $f_{u_i}^{n-1}(z_i) = f_{u_i}^{-1} \circ f_{u_i}^n(z_i) \in V$,
- $\hat{y} = f_{u_k}^{\epsilon_k} \circ \dots \circ f_{u_{i+1}}^{\epsilon_{i+1}}(z_i) \in \text{int } C(y)$.

So we have constructed a trajectory joining $z_i \in \text{int } R(x)$ to $\hat{y} \in \text{int } C(y)$ with a number of negative steps strictly less than l ; the statement follows by induction. \square

Remark 4.3. The result obtained in the previous proposition can be applied to any discrete-time system Σ that arises through the time-sampling of a continuous-time system, if the vector fields in the right-hand side of the differential equation are conservative. The latter happens for Hamiltonian systems; see for instance [10] for many examples of such Hamiltonian control systems, and the last section of [11] for conditions under which transitivity is preserved under sampling.

5. Accessibility almost everywhere. For analytic systems, we say here that a property holds for “almost all” $x \in X$ if it holds on a set which is the complement of the set of zeros of a nonzero analytic function; note that such a set is open dense and its complement has zero measure.

LEMMA 5.1. *Let Σ be an n -dimensional, discrete-time, invertible, and analytic system. Then the following are equivalent:*

- (1) Σ is transitive from almost all $x \in X$;
- (2) $\dim L(x) = n$ for almost all $x \in X$;
- (3) Σ is forward accessible from almost all $x \in X$;
- (4) $\dim L^+(x) = n$ for almost all $x \in X$.

Proof. We will show $(1) \rightarrow (2) \rightarrow (4) \rightarrow (3) \rightarrow (1)$.

$(1) \rightarrow (2)$ This is a consequence of Theorem 4 in [7].

$(2) \rightarrow (4)$ Since the system is analytic, and X is connected it will be enough to show that there is at least one x with $\dim L^+(x) = n$, because the set where this property holds is either empty or open and dense. To show that there exists such an x we will use the same procedure used in proving Lemma 4.2.

Fix any $y \in X$ for which $\dim L(y) = n$, and let $v_1, \dots, v_n \in L$ be so that

$$\{v_1(y), \dots, v_n(y)\}$$

is a basis for $L(y)$. Assume that the v_i 's involve vector fields of the form

$$Ad_0^{k_j} X_{u_j}^+,$$

with $k_j \in \mathbb{Z}$, and choose a positive integer k_0 so that $k_j + k_0 \geq 0$ for all such j . Applying the operator $Ad_0^{k_0}$ to the v_i 's, there result n linearly independent vectors in $L^+(x)$, where $x := f_0^{-k_0}(y)$. Thus $\dim L^+(x) = n$.

$(4) \rightarrow (3)$ Again by analyticity, it will be sufficient to find at least one x form which Σ is forward accessible. Choose \bar{x} regular and let k , $\mathbf{u} = (u_k, \dots, u_1)$, and \bar{z} be such that

$$\psi_{k,\bar{x}}(\mathbf{u}) = \bar{z} \quad \text{and} \quad \rho_{k,\bar{x}}(\mathbf{u}) = \bar{\rho}_{\bar{x}}.$$

Let W be some neighbourhood of \bar{x} so that

$$\rho_{k,x}(\mathbf{u}) \geq \rho_{k,\bar{x}}(\mathbf{u}) = \bar{\rho}_{\bar{x}}$$

for each $x \in W$. As \bar{x} is regular

$$\bar{\rho}_{\bar{x}} = \bar{\rho}_x \geq \rho_{k,x}(\mathbf{u}),$$

so there is equality, $\rho_{k,x}(\mathbf{u}) = \rho_{k,\bar{x}}(\mathbf{u})$. Define

$$U = f_{\mathbf{u}}(W);$$

since $f_{\mathbf{u}}$ is a diffeomorphism, U is open. Moreover, by maximality of the rank, we have

$$U \subseteq \tilde{R}^k(W).$$

Since $\dim L^+(x) = n$ for almost all x , we can choose some $z \in U$ for which $\dim L^+(z) = n$. Let

$$y := f_{\mathbf{u}}^{-1}(z) \in W.$$

Note that then $z \in \tilde{R}^k(y)$ and $\dim L^+(z) = n$.

We can conclude forward accessibility from y by (4).

(3) \rightarrow (1) This is clear. \square

Remarks 5.1. (1) Since Σ is analytic, in each of the previous statements we can substitute “there exists $x \in X$ ” instead of “for almost all $x \in X$.”

(2) Note that, in general, the open dense sets in which the previous statements hold are *not* the same, except for those in parts (1) and (2). In particular, if we denote

$$B := \{x \mid \dim L(x) < n\},$$

we have

- $B = \{x \mid x \text{ is not transitive}\};$
- $B \subseteq B_L^+ \subseteq B^+;$

and the previous inclusions can be proper. For example, for the system described in Example 6.1 below we have

$$\begin{aligned} B &= \emptyset \\ B_L^+ &= \{ (k, y) \mid k \geq 1, k \in \mathbb{Z}, -k \leq y \leq k \} \\ B^+ &= \{ (k, y) \mid k \geq 0, k \in \mathbb{Z}, -k \leq y \leq k \} = B_L^+ \cup \{(0, 0)\}. \end{aligned}$$

(3) Let L^- be the Lie algebras defined in the same way as L^+ , but using the vector fields $X_{u,i}^-$ instead of $X_{u,i}^+$, and $k \leq 0$ instead of $k \geq 0$. Given this definition, the conclusions of Lemma 5.1 hold substituting (3) and (4) with the following properties:

- (3') Σ is backward accessible from almost all $x \in X$.
- (4') $\dim L^-(x) = n$ for almost all $x \in X$.

5.1. Nonconnected orbits. Given any system Σ , its state space can be partitioned into invariant submanifolds, the orbits. Since the system restricted to each orbit is transitive, we would like to conclude that relative to each orbit there is forward accessibility from almost every state. Unfortunately, this conclusion is false in general (see Example 5.1 below), because orbits are in general not connected. We can prove this fact, however, in the particular case of orbits with at most finitely many connected components, as follows from the next result.

PROPOSITION 5.2. *Let Σ be an n -dimensional, discrete-time, invertible and analytic system, and assume that the state space X has finitely many connected components. If Σ is transitive then it is forward accessible from almost all $x \in X$.*

Proof. Partition $X = \bigcup_{i=1}^l X_i$ into disjoint nonempty open connected subsets. Note that, if $x \in X_i$ and $f(x, u) \in X_j$, then since $X_i \times U$ is connected we have that

$$(6) \quad f(X_i \times U) \subseteq X_j,$$

by continuity of f . Then for each i there is some $j(i)$ so that

$$f_u(X_i) \subseteq X_{j(i)} \quad i = 1, \dots, l,$$

for every $u \in U$.

Fix now any $u \in U$. Since $f_u(X) = X$, necessarily $\bigcup_{i=1}^l X_{j(i)} = X$. As f_u is a diffeomorphism of X , the $X_{j(i)}$ are all distinct and $f_u(X_i) = X_{j(i)}$. Since Σ is transitive, we can conclude that for any $p = 1, \dots, l-1$, denoting by

$$f_u^p = \underbrace{f_u, \dots, u}_{p\text{-times}},$$

the following holds:

$$(7) \quad f_u^p(X_i) \neq X_i \quad \forall i = 1, \dots, l.$$

If this were not the case and there exists such p and i , then applying (6) p -times we would have

$$f_{u_1, \dots, u_p}(X_i) = X_i$$

for all $(u_1, \dots, u_p) \in U^p$. Thus the set

$$\bigcup_{j=0}^{p-1} f_u^j(X_i)$$

will be an invariant set different from X , which contradicts the assumption that Σ is an orbit. Moreover, from (7), since l is finite, we can conclude that

$$(8) \quad f_{u_1, \dots, u_l}(X_i) = X_i \quad \forall i.$$

By repeating the arguments used in the proof of the Lemma 5.1 ((2) \rightarrow (4)) we conclude that there exists $x \in X$ such that $\dim L^+(x) = n$. Assume that $x \in X_i$. Since X_i is connected we have

$$\dim L^+(y) = n \quad \text{from almost all } y \in X_i.$$

Choose $\bar{x} \in X_i$, \bar{x} regular and let k , $\mathbf{u} = (u_k, \dots, u_1)$, and \bar{z} be such that

$$\psi_{k,\bar{x}}(\mathbf{u}) = \bar{z} \quad \text{and} \quad \rho_{k,\bar{x}}(\mathbf{u}) = \bar{\rho}_{\bar{x}}.$$

By inequality (2) we can assume that k is a multiple of l . Thus, by (7), we get that $\bar{z} \in X_i$. Now, we can repeat the arguments used in the proof of the Lemma 5.1 ((4) \rightarrow (3)) and conclude that Σ is forward accessible from almost all $x \in X_i$.

To conclude that Σ is forward accessible from almost all $x \in X$ it is enough to note that, for any $j \neq i$, (7) implies that there exists p such that

$$f_{u_1, \dots, u_p}(X_j) = X_i. \quad \square$$

Example 5.1. Consider the following analytic system, with $X = \mathbb{R}^2$, $U = \mathbb{R}$, and equations:

$$\begin{aligned} x^+ &= x + 1, \\ y^+ &= y + uh(x), \end{aligned}$$

where $h(x)$ is any analytic function whose zeros are exactly at the positive integers $\{1, 2, 3, \dots\}$. This system is easily seen to be invertible. Let $z_0 = (0, 0)$. Then it is easy to verify that the orbit $O(z_0)$ is as follows:

$$O(z_0) = \bigcup_{i \in \mathbb{Z}} R_i,$$

$$R_i = \{ (i, y) \mid y \in \mathbb{R} \}.$$

If we restrict the system to this orbit, the restricted system is not forward accessible from any the points in R_i , for each $i = 1, 2, 3, \dots$. This is because there it holds that $h(x) = 0$, so z^+ and z must have the same y -coordinate.

6. Low-dimensional cases. In this section we make some remarks about one- and two-dimensional systems.

6.1. Dimension one. There we consider systems for which the state space X is of dimension one. The pointwise versions of [7, Thm. 3] hold for these systems as follows.

LEMMA 6.1. *Let Σ be as above, and pick $x \in X$. Then*

1. *if Σ is smooth then*
- Σ *is forward accessible from x if and only if $\dim \Gamma^+(x) = 1$;*
2. *if Σ is analytic and U is connected then*
- Σ *is forward accessible from x if and only if $\dim L^+(x) = 1$.*

Proof. (1) The necessary part follows from part (a) of Theorem 6 in [7], so we will prove sufficiency. If Σ is not forward accessible from x then $f(x, u)$ must be independent of u . Moreover if $y = f_{u_k, \dots, u_1}(x)$, since Σ is also not forward accessible from y , also

$$f(y, u) = f(f_{u_k, \dots, u_1}(x), u)$$

must be independent of u . Thus

$$Ad_{u_k, \dots, u_1} X_{u_0}^+(x) = \frac{\partial}{\partial v} \bigg|_{v=0} f_{u_k, \dots, u_1}^{-1} \circ f_{u_0}^{-1} \circ f_{u_0+v} \circ f_{u_k, \dots, u_1}(x) = 0,$$

which implies $\dim \Gamma^+(x) = 0$.

(2) The necessary part follows from part (b) of Theorem 6 in [7]. Sufficiency is a consequence of (1), since $L^+(x) \subseteq \Gamma^+(x)$. \square

LEMMA 6.2. *Let Σ be a one-dimensional, discrete-time, invertible system, and pick any $x \in X$ so that Σ is transitive from x . Then, either Σ is forward accessible from x or Σ is backward accessible from x .*

Proof. Suppose that neither conclusion holds.

We claim that, for each $u \in U$, Σ is not forward nor backward accessible from $y = f_u(x)$. Since x is not forward accessible, $f(x, u)$ is independent of u . Thus $y = f_u(x)$ for all $u \in U$, so also

$$f_u^{-1}(y) = x \quad \text{for all } u \in U.$$

It follows that $C^1(y) = x$, which implies that

$$C^k(y) = C^{k-1}(x) \quad \text{for all } k \geq 1.$$

Thus if Σ would be backward accessible from y also Σ would be backward accessible from x . Clearly, forward accessibility from y would imply forward accessibility from x (in any dimension). So the claim is proved.

With the same arguments we can prove that Σ is not forward nor backward accessible from $z = f_u^{-1}(x)$ for all $u \in U$.

Now we want to prove that $\dim \Gamma(x) = 0$, which implies that Σ is not transitive from x . In order to do that, we will show that

$$Ad_{u_k, \dots, u_1}^{\epsilon_k, \dots, \epsilon_1} X_{u_0}^\sigma(x) = 0$$

for all $k \geq 0$, (u_k, \dots, u_1) , $\epsilon_i = 1$ or -1 , $\sigma = 1$ or -1 , and for all x which are neither forward nor backward accessible.

We will use induction on k . Take first $k = 0$.

- If $\sigma = 1$

$$X_{u_0}^+(x) = \left. \frac{\partial}{\partial v} \right|_{v=0} f_{u_0}^{-1} \circ f_{u_0+v}(x) = 0$$

since $f(x, \cdot)$ is independent of u (Σ is not forward accessible from x).

- If $\sigma = -1$

$$X_{u_0}^-(x) = \left. \frac{\partial}{\partial v} \right|_{v=0} f_{u_0} \circ f_{u_0+v}^{-1}(x) = 0$$

since $f^{-1}(x, \cdot)$ is independent of u (Σ is not backward accessible from x).

Take now any $k > 0$ and note that

$$Ad_{u_k, \dots, u_1}^{\epsilon_k, \dots, \epsilon_1} X_{u_0}^\sigma(x) = (df_{u_k}^{\epsilon_k}(x))^{-1} Ad_{u_{k-1}, \dots, u_1}^{\epsilon_{k-1}, \dots, \epsilon_1} X_{u_0}^\sigma(f_{u_k}^{\epsilon_k}(x)).$$

From the first part of the proof, we have that Σ is also neither forward nor backward accessible from $f_{u_k}^{\epsilon_k}(x)$, so, by inductive assumption, this last vector is zero. \square

Remark 6.1. A consequence of the two previous lemmas is that, for each x :

1. $\dim L(x) = 1$ if and only if $\dim L^+(x) = 1$ or $\dim L^-(x) = 1$.
2. $\dim \Gamma(x) = 1$ if and only if $\dim \Gamma^+(x) = 1$ or $\dim \Gamma^-(x) = 1$.

The result in Lemma 6.2 is true only pointwise. In fact we can find a one-dimensional, analytic system Σ that is transitive but is neither forward nor backward accessible. One example of such a system is as follows.

Consider the following system:

$$(9) \quad x^+ = 1 + x + \frac{u}{2}[g(x) + g(x-1)]$$

with $X = \mathbb{R}$, $U = [-1, 1]$, and where $g(x)$ is the following function:

$$(10) \quad g(x) = \frac{\sin(\pi x)}{\pi x}.$$

It is easy to verify that $|g'(x)| \leq 1$ for all $x \in \mathbb{R}$. Moreover, $g(x) = 0$ if and only if $x \in \mathbb{Z} \setminus \{0\}$. Since $|g'(x)| \leq 1$, this system is invertible. Moreover the following properties hold and are easily verified:

1. Σ is transitive;
2. if $x = 2, 3, \dots$ then Σ is backward accessible but not forward accessible from x ;
3. if $x = -1, -2, -3, \dots$ then Σ is forward accessible but not backward accessible from x .

6.2. Dimension two. We now show that both the results in Lemmas 6.1 and 6.2 are false if the dimension of the state space X is greater than one, even if the system is invertible, analytic and with a connected control space U .

The following example illustrates these facts.

Example 6.1. Consider the discrete-time, analytic system with $X = \mathbb{R}^2$, $U = [-1, 1]^2$, and equations:

$$\begin{aligned} x^+ &= x + 1 + \frac{u}{2} \sin(y)g(x), \\ y^+ &= y + v, \end{aligned}$$

where $g(x)$ is the function in (10).

This system is invertible. In fact, the determinant of the Jacobian matrix of the map $f_{u,v}(x, y)$ is given by

$$1 + \frac{u}{2} \sin(y)g'(x).$$

Since $u \in [-1, 1]$, $|\sin(y)| \leq 1$ and $|g'(x)| \leq 1$,

$$|\frac{u}{2} \sin(y)g'(x)| < 1$$

so the determinant is nonzero for all x, y . Moreover it is easy to verify that for each $(u, v) \in U$, the map $f_{u,v}(\cdot, \cdot)$ is bijective.

We wish to study the behavior of this system when starting from $x = 0$, $y = 0$. Let $z_0 = (0, 0)$. We prove the following properties:

- (1) the system is not forward accessible from z_0 ;
- (2) the system is not backward accessible from z_0 ;
- (3) $\dim L^+(z_0) = 2$;
- (4) the system is transitive from z_0 .

Proof.

(1) This follows from the equality

$$R^k(z_0) = \{ (k, y) \mid -k \leq y \leq k \},$$

which holds for each $k \geq 1$ and it is clear from the equations.

(2) It will be sufficient to show that

$$(11) \quad C^k(z_0) = \{ (-k, y) \mid -k \leq y \leq k \}.$$

First note that if $(x_k, y_k) \in C^k(z_0)$ then we can write $y_k + v_1 + \dots + v_k = 0$ with all $|v_i| \leq 1$, so $|y_k| \leq k$.

To prove (11), it is now sufficient to note the following. For any fixed $u \in [-1, 1]$ and any $y \in \mathbb{R}$, the function

$$x \mapsto x + 1 + \frac{u}{2} \sin(y)g(x)$$

is invertible. Moreover

$$h^{-1}(-k+1) = -k \quad \text{for all } k \geq 1,$$

independently of u and y . Thus $x = -k$ is the only solution of $h(x) = -k+1$, for all u, y , and we have proved

$$C^k(z_0) \subseteq \{ (-k, y) \mid -k \leq y \leq k \}.$$

The other inclusion is obvious.

(3) Consider the vector fields

$$X_{(u,v),1}^+(z) = (df_{(u,v)}(z))^{-1} \frac{\partial}{\partial \epsilon} \Big|_{\epsilon=0} f_{(u+\epsilon,v)}(z)$$

and

$$X_{(u,v),2}^+(z) = (df_{(u,v)}(z))^{-1} \frac{\partial}{\partial \epsilon} \Big|_{\epsilon=0} f_{(u,v+\epsilon)}(z).$$

Fix $(u, v) = (0, 0)$. Then

$$df_{(0,0)}(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

for all $z \in \mathbb{R}^2$,

$$X_{(0,0),1}^+(z) = \begin{pmatrix} \frac{\sin(y)g(x)}{2} \\ 0 \end{pmatrix},$$

and

$$X_{(0,0),2}^+(z) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

So

$$\left[X_{(0,0),1}^+, X_{(0,0),2}^+ \right](z) = \begin{pmatrix} \frac{\cos(y)g(x)}{2} \\ 0 \end{pmatrix}.$$

In particular,

$$X_{(0,0),2}^+(z_0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and

$$\left[X_{(0,0),1}^+, X_{(0,0),2}^+ \right] (z_0) = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix},$$

so $\dim L^+(z_0) = 2$ as desired.

(4) Transitivity at z_0 is a consequence of (3) since $\dim L^+(z_0) = 2$ implies $\dim L(z_0) = 2$. \square

7. Controllability sets. The next definition is a discrete-time analogue of that in [4], except that we make the assumption of nonempty interior.

DEFINITION 7.1. A set $D \subseteq X$ is called a *precontrollability set* if

$$D \subseteq \text{clos } R(x) \quad \text{for all } x \in D$$

and $\text{int } D \neq \emptyset$. A precontrollability set which is maximal with respect to set inclusion is called a *controllability set*.

Note that if D is a precontrollability set, then in D the system Σ is “almost” controllable, in the sense that if $x, y \in D$ then from x it is possible to reach any neighbourhood of y .

LEMMA 7.1. Let $D \subseteq X$ be a controllability set. Pick any two elements \bar{x}, \bar{y} in $\text{int } D$. Then, for each sequence $(u_0, \dots, u_T) \in U^{T+1}$ such that

$$f_{u_T} \circ \dots \circ f_{u_0}(\bar{x}) = \bar{y}$$

we have that, necessarily, also

$$f_{u_k} \circ \dots \circ f_{u_0}(\bar{x}) \in \text{int } D \quad \text{for } k = 0, \dots, T-1.$$

Proof. Let $\bar{x}, \bar{y}, u_0, \dots, u_T$ be as in the statement and let E be the following set:

$$E := \{ f_{u_k} \circ \dots \circ f_{u_0}(\bar{x}), \quad k = 0, \dots, T-1 \}.$$

We will first prove that $E \subseteq D$, by showing that $D' = D \cup E$ is again a precontrollability set and using that D is maximal. For this, we must prove that

$$D' \subseteq \text{clos } R(x) \quad \text{for each } x \in D'.$$

Observe that $E \subseteq R(\bar{x}) \subseteq \text{clos } R(\bar{x})$ and $\bar{y} \in R(y) \subseteq \text{clos } R(y)$ for all $y \in E$. Thus

- $E \subseteq \text{clos } R(\bar{x}) \subseteq \text{clos } R(\text{clos } R(x)) = \text{clos } R(x) \quad \forall x \in D$;
- $D \subseteq \text{clos } R(x) \quad \forall x \in D$;
- If $y \in E$ then $D \subseteq \text{clos } R(\bar{y}) \subseteq \text{clos } R(\text{clos } R(y)) = \text{clos } R(y)$
and $E \subseteq \text{clos } R(\bar{x}) \subseteq \text{clos } R(\text{clos } R(\bar{y})) \subseteq \text{clos } R(y)$.

Thus $D \cup E = D' \subseteq \text{clos } R(x) \quad \forall x \in D'$.

So we have proved that, for any two points \bar{x}, \bar{y} in $\text{int } D$ and any trajectory joining them, all the intermediate states must be in D . We now prove that such intermediate points must be in $\text{int } D$.

Pick any $\bar{x}, \bar{y}, u_0, \dots, u_T$ as above. Let $k \in \{0, \dots, T-1\}$ and $\bar{z} = f_{u_k} \circ \dots \circ f_{u_0}(\bar{x})$. By continuity of $f_{u_0}^{-1} \circ \dots \circ f_{u_k}^{-1}$ and of $f_{u_{k+1}} \circ \dots \circ f_{u_T}$, there exists some open neighbourhood V of \bar{z} such that

$$f_{u_0}^{-1} \circ \dots \circ f_{u_k}^{-1}(V) \subseteq \text{int } D \quad \text{and} \quad f_{u_{k+1}} \circ \dots \circ f_{u_T}(V) \subseteq \text{int } D.$$

Pick any $z \in V$. For such a z ,

$$z = f_{u_k} \circ \dots \circ f_{u_0}(x)$$

for some $x \in \text{int } D$ and

$$y = f_{u_{k+1}} \circ \dots \circ f_{u_T}(z) \in \text{int } D.$$

Thus, applying the first part of the proof to x and y (rather than to \bar{x} and \bar{y}), it follows that $z \in D$. We conclude that $V \subseteq D$, so \bar{z} is in $\text{int } D$, as desired. \square

LEMMA 7.2. *Let $D \subseteq X$ be a precontrollability set. Then we have*

$$D \subseteq \text{clos } F_k(\text{int } D) \quad \text{for all } k = 0, 1, 2, \dots,$$

where

$$F_k(\text{int } D) = \bigcup_{l \geq k} R^l(\text{int } D).$$

Proof. We proceed by induction on k . The case $k = 0$ follows directly from the definition of controllability set. So let $k \geq 1$ and pick any $x \in D$.

Choose $y \in \text{int } D$, $y \neq x$. By inductive assumption there exists a sequence $y_n \rightarrow y$ with

$$y_n \in F_k(\text{int } D).$$

For \bar{n} sufficiently large, $y_{\bar{n}} \in D$ (since $y \in \text{int } D$) and $y_{\bar{n}} \neq x$ (since $y \neq x$), where each $y_{\bar{n}}$ is of the form

$$y_{\bar{n}} = \psi_{l,z}(\mathbf{u})$$

with $z \in \text{int } D$, $l \geq k$, for some $\mathbf{u} \in U^l$. Pick one such \bar{n} . Since $x \in \text{clos } R(y_{\bar{n}})$, there exist a sequence $\{t_n\}$ and a sequence $\{z_n\}$ so that

$$z_n \in R^{t_n}(y_{\bar{n}}) \quad \text{and} \quad z_n \rightarrow x.$$

Since $y_{\bar{n}} \neq x$ we can assume $t_n \geq 1$ for all n . Thus

$$z_n \in R^{l+t_n}(z) \subseteq F_{k+1}(\text{int } D),$$

which implies $x \in \text{clos } F_{k+1}(\text{int } D)$. \square

Remark 7.1. The conclusion of the previous lemma can be rephrased by saying that

$$D \subseteq \overline{\lim}_k R^k(\text{int } D),$$

where for any family of sets E_k , $\overline{\lim}_k E_k = \bigcap_{k=0}^{\infty} \bigcup_{l \geq k} E_l$.

LEMMA 7.3. *Let $D \subseteq X$ be a controllability set. Then*

$$\text{clos } D = \text{clos int } D.$$

Proof. Let $x \in D$. We only need to prove that for any neighbourhood W of x ,

$$W \cap \text{int } D \neq \emptyset.$$

Pick any such W and choose any $y \in \text{int } D$. Since $y \in \text{clos } R(x)$, we can find $z = \psi_{k,x}(\mathbf{u})$ for some $k \geq 0$ and some $\mathbf{u} \in U^k$, such that $z \in \text{int } D$. Let U_z be a neighbourhood of z contained in D . Then, by continuity, there exists a neighbourhood U_x of x such that for all $y \in U_x$, $\psi_{k,y}(\mathbf{u})$ is in U_z and so, in particular, in $\text{int } D$.

Let $W_x = U_x \cap W$. Choose $y' \in \text{int } D$. Since $x \in \text{clos } R(y')$, we can find k', \mathbf{u}' such that

$$\bar{x} := \psi_{k',y'}(\mathbf{u}') \in W_x.$$

Let $\bar{\mathbf{u}}$ be the concatenation of \mathbf{u}' and \mathbf{u} . Since $\bar{x} \in U_x$,

$$\psi_{k,\bar{x}}(\mathbf{u}) \in \text{int } D.$$

Thus

$$\psi_{k+k',y'}(\bar{\mathbf{u}}) \in \text{int } D,$$

so by Lemma (7.1), $\bar{x} \in \text{int } D$. Hence

$$W_x \cap \text{int } D \neq \emptyset,$$

so $W \cap \text{int } D \neq \emptyset$ as wanted. \square

DEFINITION 7.2. Let $x \in X$ and $S \subseteq X$. We say that x is *forward accessible in S* (respectively, *backward accessible in S*) if

$$\text{int}(R(x) \cap S) \neq \emptyset$$

(respectively, $\text{int}(C(x) \cap S) \neq \emptyset$).

If we simply say that x is forward (backward) accessible, we mean forward (backward) accessible in X .

LEMMA 7.4. *Let $S \subseteq X$ and define*

$$S_f = \{ x \in M \mid x \text{ is forward accessible in } S \},$$

then S_f is open.

Proof. If $S_f = \emptyset$, then it is trivially open; thus assume $S_f \neq \emptyset$. Pick any $x \in S_f$.

By assumption there exists $W \subseteq S$ open such that $W \subseteq R(x)$; therefore there exists k such that $W \cap R^k(x)$ has nonzero measure. Let

$$U_W^k = \{ \mathbf{u} \mid \mathbf{u} \in U^k, \text{ and } \psi_{k,x}(\mathbf{u}) \in W \},$$

then U_W^k is open and the image of

$$\psi_{k,x}|_{U_W^k}$$

has nonzero measure. It follows, by Sard's theorem, that there exists $\mathbf{u} \in U^k$ such that $\rho_{k,x}(\mathbf{u}) = n$. We may assume, without loss of generality, that $\mathbf{u} \in \text{int } U^k$.

Now pick any neighbourhood V of x such that $\psi_{k,V}(\mathbf{u}) \subseteq W$ and still $\rho_{k,y}(\mathbf{u}) = n$ for all $y \in V$. By the implicit mapping theorem, $V \subseteq S_f$; therefore S_f is open. \square

We assume from now on that the system Σ to be *analytic* and *transitive*. In this case, we can conclude the following important property of precontrollability sets.

THEOREM 7.5. *Let $D \subseteq X$ be a precontrollability set. Then every point of D is forward accessible in D .*

Proof. Since Σ is transitive and analytic, by Lemma 5.1 we have that there exists an open dense set of points for which $\dim L^+(x) = n$. If we intersect this set with $\text{int } D$ then this intersection, which we denote by W , is open.

Pick any $x \in D$, and y in W , with $x \neq y$. Now, we will construct a sequence of elements y_n such that

$$y_n \rightarrow y, \quad y_n \in R^{k_n}(x) \quad \text{and} \quad k_n \rightarrow \infty \quad \text{as} \quad n \rightarrow \infty.$$

Then, using Lemma 4.1 and its successive remarks, we can conclude that from x it is possible to reach an open set within any neighbourhood of y , i.e., since $y \in \text{int } D$, x is forward accessible in D , as desired.

To construct the y_n 's we proceed as follows. Let's denote by W_n a neighbourhood of y . Since $y \in \text{clos } R(x)$, we can find $y_1 \in W_1$, $y_1 \neq y$, and $y_1 \in R^{k_1}(x)$ where (since $y \neq x$) we can assume $k_1 \geq 1$.

Now we proceed by induction. Suppose that we have found y_1, \dots, y_n such that

$$y_i \in W_i, \quad y_i \neq y, \quad y_i \in R^{k_i}(x) \quad \text{and} \quad k_i \geq i \quad \text{for} \quad i = 1, \dots, n.$$

Since $y \in \text{clos } R(y_n)$ we can find $y_{n+1} \in W_{n+1}$ such that

$$y_{n+1} \neq y, \quad y_{n+1} \in R^l(y_n)$$

with $l \geq 1$. Since $y_n \in R^{k_n}(x)$,

$$y_{n+1} \in R^{k_n+l}(x)$$

and $k_{n+1} = k_n + l \geq n + 1$. Thus $k_n \rightarrow \infty$ as $n \rightarrow \infty$; moreover, we can choose the W_i 's in such a way that $y_n \rightarrow y$. \square

The definition of precontrollability set is not reversible in time, so we cannot conclude backward accessibility from every point. However, the next result provides backward accessibility from a dense subset.

PROPOSITION 7.6. *Let $D \subseteq X$ be a controllability set. Then there exists some (necessarily nonempty) open subset $E \subseteq D$ such that*

- (1) $\text{clos } E = \text{clos } D$;
- (2) if $y \in E$ then y is backward accessible in D .

Proof. Since Σ is transitive and analytic, by Lemma 5.1 applied to the "inverse" system

$$x(t+1) = f_u^-(x(t)),$$

we know that there exists an open dense set from which we have backward accessibility. Moreover, there exists some integer k_0 such that the set G of states $x \in X$ for which $\text{int } C^{k_0}(x) \neq \emptyset$ and

$$\text{clos int } C^k(x) = \text{clos } C^k(x)$$

for all $k \geq k_0$ is itself open dense (Remark 3.2). Consider first the open set

$$E' = \text{int } D \cap G.$$

We claim that E' is open and

$$\text{clos } E' = \text{clos int } D.$$

To show this, it is enough to establish that $\text{int } D \subseteq \text{clos } E'$. So take any $x \in \text{int } D$. By density of G , there exists some sequence $\{y_n\}$ with $y_n \in G$ for all n , $y_n \rightarrow x$. Thus

$$y_n \in \text{int } D \cap G$$

for all large enough n , and this shows that $x \in \text{clos } E'$. Finally, let

$$E = E' \cap F_{k_0}(\text{int } D),$$

where $F_{k_0}(\text{int } D)$ is defined as in Lemma (7.2). Then E is also open, since $F_k(\text{int } D)$ is open for any k . Moreover, using the result in Lemma (7.2) (i.e., $D \subseteq \text{clos } F_{k_0}(\text{int } D)$) and the same arguments used before we have

$$\text{clos } E = \text{clos } E' = \text{clos int } D.$$

Thus, by Lemma (7.3),

$$\text{clos } E = \text{clos } D.$$

So E satisfies property (1). We prove next that it also satisfies (2).

Pick $y \in E$. Since $y \in F_{k_0}(\text{int } D)$ then there exists $x \in \text{int } D$ so that $y \in R^k(x)$ for some $k \geq k_0$. This means that $x \in C^k(y)$. Since $y \in G$,

$$x \in \text{clos int } C^k(y).$$

Thus, since $x \in \text{int } D$, we can find some $z \in \text{int } D \cap \text{int } C^k(y)$, which means that y is backward accessible in D . Thus (2) is proved. \square

LEMMA 7.7. *Let $D \subseteq X$ be a controllability set and let E be any set as in the conclusion of the previous proposition. Then*

$$E \subseteq R(x) \quad \text{for each } x \in D.$$

Proof. Take any $y \in E$ and $x \in D$. By the previous proposition, there exists some nonempty open set $W \subseteq D \cap C(y)$. Choose any $z \in W$. Since D is a controllability set, $z \in \text{clos } R(x)$, so there exists also $\tilde{z} \in R(x) \cap W$. Thus $\tilde{z} \in R(x)$ and $y \in R(\tilde{z})$ (since $\tilde{z} \in C(y)$) imply $y \in R(x)$. \square

DEFINITION 7.3. For any set $S \subseteq X$, define

$$\text{Core}(S) := \{ x \in \text{int } S \mid x \text{ is forward and backward accessible in } S \}.$$

Using Lemma 7.4 twice (once for Σ and another time for the “inverse” system $x(t+1) = f_u^-(x(t))$), we can conclude the following.

LEMMA 7.8. *For any subset $S \subseteq X$, $\text{Core}(S)$ is open.*

For a controllability set D , we proved (see results in Theorem 7.5 and Proposition (7.6)) that $\text{Core}(D) \supseteq E$ for some $E \subseteq D$ such that $\text{clos } E = \text{clos } D$. Thus we have

$$(12) \quad \boxed{\text{clos Core}(D) = \text{clos } D \quad \text{for a controllability set } D.}$$

Moreover, the result in Lemma (7.7) can be rephrased as follows.

PROPOSITION 7.9. *If D is a controllability set, and $E = \text{Core}(D)$, then $E \subseteq R(x)$ for all $x \in D$.*

If D is a controllability set, then, by the previous results, $\text{Core}(D)$ is a dense subset of D in which we have exact controllability. Note that if Σ was a continuous time system then $\text{Core}(D)$ would have been equal to $\text{int } D$. However for discrete-time systems there are controllability sets D for which $\text{Core}(D)$ is strictly contained in $\text{int } D$, as it is shown in the next example.

Example 7.1. Let us consider the discrete-time, analytic system with $X = \mathbb{R}^2$, $U = [-1, 1]^2$, and equations

$$\begin{aligned}x^+ &= x + 1 + uy, \\y^+ &= y + \frac{v}{2}g(x),\end{aligned}$$

where $g(x)$ is the function in (10).

This system is invertible. In fact the determinant of the Jacobian matrix of the map $f_{u,v}(x, y)$ is given by

$$1 - \frac{uv}{2}g'(x).$$

Since $u, v \in [-1, 1]$, and $|g'(x)| \leq 1$,

$$\left| \frac{uv}{2}g'(x) \right| \leq \frac{1}{2}$$

so the determinant is nonzero for all x, y . Moreover it is easy to verify that for each $(u, v) \in U$, the map $f_{u,v}(\cdot, \cdot)$ is bijective. It is also easy to prove that this system is transitive.

For this system we can see that for all $k \in \mathbb{N}$ with $k \geq 1$ the following hold:

1. the points of the type $(-k, 0)$ are not backward accessible;
2. the points of the type $(k, 0)$ are not forward accessible.

Let

$$B = \{ (k, 0) \mid k \in \mathbb{N}, \ k \geq 1 \}.$$

Next we want to show that $D = \mathbb{R}^2 \setminus B$ is a controllability set.

Note that D is certainly maximal; in fact, no points in B could belong to a controllability set, since they are not forward accessible. To prove that D satisfies

$$(13) \quad D \subseteq \bar{R}(\xi) \quad \text{for all } \xi \in D$$

we will prove the following:

$$(14) \quad \mathbb{R}^2 \setminus \{ (k, y) \mid k \in \mathbb{Z}, \ y \in \mathbb{R} \} \subseteq R(\xi),$$

which, by taking the closure in both sides, implies (13). Let $F = \{ (k, y) \mid k \in \mathbb{Z}, \ y \in \mathbb{R} \}$.

First we note that, since $|\sin(\pi(x+1))| = |\sin(\pi x)|$, if we apply to any (x, y) a control sequence of the following form:

$$(15) \quad u_l = 0, \quad v_l = \text{sign}(g(x+l-1)),$$

then, after k steps, we will reach the following point:

$$x_k = x + k$$

$$y_k = y + \frac{|\sin(\pi x)|}{2\pi} \sum_{l=0}^{k-1} \frac{1}{|x+l|}.$$

Using this fact and the divergence of the series $\sum_n 1/n$ we will prove (14).

Fix $(\bar{x}, \bar{y}) \in D$ and $(\tilde{x}, \tilde{y}) \in \mathbb{R}^2 \setminus F$. Note that, since $(\bar{x}, \bar{y}) \notin B$, it is not restrictive to assume

$$g(\bar{x}) \neq 0 \quad \text{and} \quad \bar{y} \neq 0.$$

First we choose u_l, v_l as in (15). Since $g(\bar{x}) \neq 0$ there exists k such that $y_k > 1$. Next we apply a control sequence with all $v_l = 0$ so as to reach a state (x', y') of the type

$$x' = \tilde{x} - n \quad \text{and} \quad y' = y_k,$$

where n is a positive integer that will be chosen later. Note that we can assume $\tilde{y} < y'$.

Now we want to find a sequence of controls $(0, v_l)$ such that we get the state (\tilde{x}, \tilde{y}) in exactly n steps. It is clear that this is possible if and only if

$$(16) \quad y' - \frac{|\sin(\pi \tilde{x})|}{2\pi} \sum_{l=0}^{n-1} \frac{1}{|\tilde{x} - n + l|} \leq \tilde{y}.$$

So we just have to choose n large enough such that (16) is satisfied. This is possible since $\sin(\pi \tilde{x}) \neq 0$ and

$$\sum_{l=0}^{n-1} \frac{1}{|\tilde{x} - n + l|} = \sum_{m=1}^n \frac{1}{|\tilde{x} - m|}$$

is divergent. Thus D is a controllability set.

Note that, for this controllability set D , $\text{Core}(D)$ is strictly contained in $D = \text{int } D$. In fact, none of the points of the type $(-k, 0)$ with k a strictly positive integer belongs to $\text{Core}(D)$.

REFERENCES

- [1] F. ALBERTINI AND E. D. SONTAG, *Some connections between chaotic dynamical systems and control systems*, in Proc. European Control Conference, Vol. 1, Grenoble, July 1991, pp. 158–163.
- [2] ———, *Further remarks on controllability properties for discrete-time nonlinear systems*, Dynamics and Control, to appear, 1994.
- [3] J. P. BARBOT, *A forward accessibility algorithm for nonlinear discrete-time systems*, in Proc. 9th Conf. Anal. and Optim., Lecture Notes in Control and Information Science, No. 144, Springer-Verlag, pp. 314–323, 1990.
- [4] F. COLONIUS AND W. KLIEMANN, *Some Aspects of Control Systems as Dynamical Systems*, Report No. 223, Inst. Math., Univ. Augsburg, 1990.
- [5] M. FLIESS AND D. NORMAND-CYROT, *A group-theoretic approach to discrete-time nonlinear controllability*, Proc. IEEE Conf. Dec. Control, San Diego, Dec. 1981.
- [6] B. JAKUBCZYK AND D. NORMAND-CYROT, *Orbites de pseudo groupes de difféomorphismes et commandabilité des systèmes non linéaires en temps discret*, C. R. Acad. Sciences de Paris, 298-I (1984), pp. 257–260.

- [7] B. JAKUBCZYK AND E. D. SONTAG, *Controllability of nonlinear discrete-time systems: A Lie-algebraic approach*, SIAM J. Control and Optim., 28 (1990), pp. 1–33.
- [8] C. LOBRY, *Controllability of nonlinear systems on compact manifolds*, SIAM J. Control, 12 (1974), pp. 1–4.
- [9] A. MOKKADEM, *Orbites de semi-groupes de morphismes réguliers et systèmes non linéaires en temps discret*, Forum Math, Vol 1, pp. 359–376, 1989.
- [10] H. NIJMEIJER AND A. V. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
- [11] E. D. SONTAG, *Integrability of certain distributions associated with actions on manifolds and applications to control problems*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, NY 1990, pp. 81–131.
- [12] ———, *Mathematical Control Theory, Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.
- [13] H. J. SUSSMANN, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 172–188.

INFINITE DETERMINISTIC GRAPHICAL GAMES*

V. J. BASTON[†] AND F. A. BOSTOCK[†]

Abstract. A deterministic graphical game is a two-person zero-sum game with perfect information played on a directed graph. Nodes with no successor are called terminal nodes and have a payoff associated with them. The other nodes are labelled to indicate which of the players chooses the successor node. Play starts at a specific node and only stops when a terminal node is reached at which point player 1 obtains the payoff corresponding to that node; if play never ends the payoff is zero. The paper shows that such games have a solution even when the graph has an infinite number of nodes.

Key words. recursive game, two-person zero-sum game, game with perfect information, graphical game

AMS subject classifications. 90D20, 90D05

1. Introduction. Two-person zero-sum games with perfect information have attracted considerable attention over the years. They were defined by von Neumann and Morgenstern [10], who proved that they have a value and that the players in them have pure strategies that are optimal. Gale and Stewart [5] considered infinite games with perfect information in which a play of the game gives rise to a sequence of digits that determines a real number x leading to a payoff $f(x)$. They demonstrated that not all such games have a solution and found sufficient conditions for them to do so. Davis [2] and Mycielski [6] obtained further results on these games while Oxtoby [8] investigated a very general game of a similar type. Ehrenfeucht and Mycielski [3] studied games with perfect information in which the two players move alternately along the edges of a finite directed graph having weights attached to its edges. More recently, Washburn [11] considered a class of perfect information games played on a finite graph and his results have been supplemented by Baston and Bostock [1]. Washburn called his games deterministic graphical games and defined them as follows.

A *deterministic graphical game* is a two-person zero-sum game played on a directed graph with $n > 0$ nodes. Nodes with no successor are called terminal nodes and have a payoff to player 1 associated with them. The other nodes are called continuing nodes and are labelled to indicate which player chooses the successor node. Play starts at a specific node and only stops when a terminal node is reached at which point player 1 obtains the payoff corresponding to that node; if play never ends, the payoff is by convention zero.

Washburn pointed out that special cases of these games had previously occurred in the literature but, that up to then, the class had remained nameless. He also commented that they are special cases of recursive games with the consequence that every deterministic graphical game has a value and, what is more, attention can be restricted to stationary strategies for the players. In this paper we consider the corresponding game in which the number of nodes can be infinite; since it is still an open question whether all countably infinite recursive games have a solution, the position

*Received by the editors October 16, 1991; accepted for publication (in revised form) July 15, 1992.

[†]Faculty of Mathematical Studies, University of Southampton, Southampton SO9 5NH, United Kingdom.

in this case is not so straightforward. However, we show that even uncountably infinite deterministic graphical games always have a value and that the players have pure stationary strategies that are essentially ϵ -optimal (with an appropriate interpretation when a component game has an infinite value). In recursive games, knowing the starting component is regarded as part of the history and, in this context, the strategies we present are very simple history remembering (namely semistationary) ones. An example is provided to show that infinite recursive games with perfect information need not have a solution if the players are restricted to Markov strategies which were the strategies used by Everett in his paper [4] introducing finite recursive games.

2. Notation and preliminary ideas. Let W be the set of nodes that we consider to be well-ordered; the well-ordering is used when we select certain elements in the proof of Theorem 3.1. Let T be the set of terminal nodes, I the set of continuing nodes at which player 1 chooses the successor, and J the set of continuing nodes at which player 2 chooses the successor. For $k \in W$, let

$$U_k = \begin{cases} \{i \in W : i \text{ is a successor of } k\} & \text{if } k \in I; \\ \{1\} & \text{if } k \in W \setminus I, \end{cases}$$

and

$$V_k = \begin{cases} \{i \in W : i \text{ is a successor of } k\} & \text{if } k \in J; \\ \{1\} & \text{if } k \in W \setminus J. \end{cases}$$

Note that U_k and V_k are effectively the control sets of players 1 and 2 at node k ; since player 1 has no control over which successor is chosen at nodes k in $W \setminus I$, we have, in such cases, defined U_k arbitrarily as $\{1\}$. Analogous comments apply to the V_k .

For $k \in W$, $u \in U_k$, $v \in V_k$ let

$$f(k, u, v) = \begin{cases} u & \text{if } k \in I; \\ v & \text{if } k \in J; \\ k & \text{if } k \in T, \end{cases}$$

and

$$c(k, u, v, f(k, u, v)) = \begin{cases} P(f(k, u, v)) & \text{if } k \notin T, f(k, u, v) \in T; \\ 0 & \text{otherwise,} \end{cases}$$

where, for $w \in T$, $P(w)$ is the payoff to player 1 at w .

We now define our strategy spaces for the players. In a deterministic graphical game each player knows the node at which the game starts. However, in a multistage game, it is usual for a strategy to provide instructions for all starting nodes and this is the approach we will adopt. A *history* h_k at stage $k \geq 0$ is defined as a sequence

$$(1) \quad (x_0; u_0, v_0), (x_1; u_1, v_1), \dots, (x_{k-1}; u_{k-1}, v_{k-1}), x_k,$$

where $x_i \in W$, $u_i \in U_{x_i}$, and $v_i \in V_{x_i}$. Note that a history at stage $k = 0$ is not empty but comprises of the starting node. Our definition of history corresponds to that of, for instance, Nowak and Raghavan [7] in stochastic games. Let \mathcal{H}_k be the set of all histories at stage $k \geq 0$. A *history remembering strategy* F for player 1 is a sequence F^0, F^1, \dots , where F^k is a function from \mathcal{H}_k to $\bigcup_{i \in W} U_i$, where $F^k(h_k) \in U_{x_k}$ when h_k is given by (1). A history remembering strategy for player 2 is defined similarly.

If, for all k and for all h_k of the form (1), $F^k(h_k)$ depends solely on the starting node x_0 and the node x_k that play has reached, F is called *semistationary*. When, in addition, $F^k(h_k)$ is independent of the starting node x_0 for all k the

strategy is said to be *stationary*. Given a starting node x_0 and strategies F and G for player 1 and player 2, respectively, it is easily seen that an infinite sequence $(x_0, u_0, v_0, x_1), (x_1, u_1, v_1, x_2), \dots$ is defined recursively where $f(x_r, u_r, v_r) = x_{r+1}$; the payoff $C(x_0, F, G)$ is now given by $C(x_0, F, G) = \sum_{k=0}^{\infty} c(x_k, u_k, v_k, x_{k+1})$. Player 1 seeks to maximize $C(x_0, F, G)$ and player 2 seeks to minimize it.

A deterministic graphical game has been modelled as a system evolving in W according to $x_{t+1} = f(x_t, u_t, v_t)$ starting at a given node $x_0 \in W \setminus T$ at stage $t = 0$; we are ignoring the trivial case when play starts at a terminal node. The ϵ -optimal strategies we will give for the players are semistationary strategies. Note that, in the system, play effectively terminates if and when play first enters a terminal node since the only possible nonzero term of the infinite sum is one that corresponds to play first entering a terminal node.

For the example in §4 it is convenient to use the notation of recursive games (see [4] or, on a more elementary level, [9]). Note that our system can be put as a recursive game $\Gamma = (\Gamma_i)_{i \in W}$, where Γ_i has matrix $M_i : U_i \times V_i \rightarrow W$ given for $u \in U_i, v \in V_i$ by

$$M_i(u, v) = \begin{cases} \Gamma_{f(i, u, v)} & \text{if } i \notin T; \\ P(i) & \text{if } i \in T. \end{cases}$$

For all nonnegative integers n and all positive real numbers x we now define a subset $H_n(x)$ of the set of nodes W . For a positive real number x put

$$H_0(x) = \{t \in T : P(t) \geq x\}.$$

Let $n \geq 1$ be a positive integer and suppose that $H_r(x)$ has been defined for all $x > 0$ and for all nonnegative integers $r < n$. For $x > 0$ define

$$H_n(x) = \{i \in I : U_i \cap H_r(x) \text{ is nonempty for some } r < n\} \\ \cup \{i \in J : \text{for each } \xi \in V_i \text{ there is an } r < n \text{ such that } \xi \in H_r(x)\}.$$

Clearly $r < n$ implies $H_r(x) \subseteq H_n(x)$. Further, using induction, it is easy to see that $x < y$ implies $H_n(y) \subseteq H_n(x)$.

Let $H(x) = \bigcup_{n=0}^{\infty} H_n(x)$. Intuitively, if play starts at a node $i \in H(x)$, then player 1 will be able to ensure himself of at least x . Now define $p(i) = \max\{0, \sup\{x : i \in H(x)\}\}$. Note that we write $p(i) = \infty$ when $\{x : i \in H(x)\}$ is not bounded above. For $x > 0$ define $m_j(x)$ as follows; if $j \in H(x) \cap I$, then $m_j(x)$ is the least integer n such that $U_j \cap H_n(x)$ is nonempty; otherwise $m_j(x) = 0$.

In an analogous fashion, for a negative real number y , put

$$K_0(y) = \{t \in T : P(t) \leq y\}$$

and, for positive integers $n \geq 1$,

$$K_n(y) = \{i \in J : V_i \cap K_r(y) \text{ is nonempty for some } r < n\} \\ \cup \{i \in I : \text{for each } \xi \in U_i \text{ there is an } r < n \text{ such that } \xi \in K_r(y)\}.$$

Let $K(y) = \bigcup_{r=0}^{\infty} K_r(y)$ and define $n(i) = \min\{0, \inf\{y : i \in K(y)\}\}$, where possibly $n(i) = -\infty$.

3. Main result. In this section we show that an infinite deterministic graphical game has a value and obtain an expression for it in terms of the $p(i)$ and $n(i)$. In coping with the fact that the game may have an infinite value, we remind the reader that, in a game with an infinite value, a strategy X^* is said to be ϵ -optimal for player

1 if the expectation $E(X^*, Y) > 1/\epsilon$ for all strategies Y for player 2. A corresponding definition applies for player 2.

THEOREM 3.1. *The value v_i of the deterministic graphical game starting at node i is given by*

$$v_i = \begin{cases} p(i) & \text{if } p(i) \neq 0; \\ n(i) & \text{otherwise.} \end{cases}$$

Furthermore, for this particular game that starts at the node i , the players possess pure ϵ -optimal stationary strategies for all positive numbers ϵ .

Proof. Let $\epsilon > 0$. For the game starting at node i we now define a pure stationary strategy $X^*(i)$ for player 1.

If $p(i) \neq 0$ then $p(i) > 0$ or $p(i) = \infty$ by the definition of $p(i)$. Choose a positive number x_i such that $p(i) - \epsilon < x_i < p(i)$ if $p(i) \neq \infty$ and $x_i > 1/\epsilon$ if $p(i) = \infty$. It is easy to see that $i \in H(x_i)$. Now let $X^*(i)$ be the stationary strategy which, for each $j \in I$, selects the first element in U_j belonging to $H_{m_j(x_i)}(x_i)$ if $j \in H(x_i)$, and the first element in U_j otherwise.

Now suppose $p(i) = 0$. If $n(i) = -\infty$ let $X^*(i)$ be the strategy that chooses the first element in U_j for every j satisfying $j \in I$ (in maintaining $-\infty$ player 1 can be as careless as he likes!). If $n(i) \neq -\infty$ then $n(i) \leq 0$ so choose $y_i < 0$ such that $n(i) - \epsilon < y_i < n(i)$. Thus $i \notin K(y_i)$. Now, for each $j \in I$ satisfying $j \notin K(y_i)$, there is an element in U_j not belonging to $K_r(y_i)$ for any positive integer r (and therefore not in $K(y_i)$) and so we can let $X^*(i)$ select the first such element in U_j . Otherwise let $X^*(i)$ select the first element in U_j .

We now show that $X^*(i)$ is ϵ -optimal. Suppose player 1 uses strategy $X^*(i)$, player 2 the strategy Y .

(I) Assume $p(i) \neq 0$ then $i \in H(x_i)$. Suppose at stage $t \geq 1$ under $X^*(i)$ and Y that either play has reached a terminal node with a payoff to player 1 of at least x_i or play has been sent through the nodes $i = i(1), i(2), \dots, i(t)$ where $i(j) \in H_{n(j)}(x_i)$ for $j = 1, 2, \dots, t$ and $n(1) > n(2) > \dots > n(t)$. Suppose the latter holds; note that this is the position for $t = 1$ if i is not a terminal node.

If $i(t) \in I$ then under $X^*(i)$ player 1 chooses the first element in $U_{i(t)}$ belonging to $H_{m_{i(t)}(x_i)}(x_i)$. Thus either play reaches a terminal node giving player 1 a payoff of at least x_i or play is sent to a node $i(t+1) \in H_{n(t+1)}(x_i)$ where $n(t+1) < n(t)$.

If $i(t) \in J$ then the fact that $i(t) \in H_{n(t)}(x_i)$ implies that, whatever the strategy Y , play has to be sent to a node that is in an $H_{n(t+1)}(x_i)$, where $n(t+1) < n(t)$. Thus again either play reaches a terminal node giving player 1 a payoff of at least x_i or play is sent to a node $i(t+1) \in H_{n(t+1)}(x_i)$, where $n(t+1) < n(t)$.

Hence at stage $t+1$ we have similar conditions to those at stage t and the process can be repeated if play has not reached a terminal node. The process must clearly reach a terminal node after at most $n(1)$ steps and so player 1 is ensured of a payoff of at least x_i .

(II) Now assume $p(i) = 0$ and $n(i) \neq -\infty$ then $i \notin K(y_i)$. Suppose at stage $t \geq 1$ under $X^*(i)$ and Y that either play has reached a terminal node with a payoff to player 1 of at least y_i or play has been sent through the nodes $i = i(1), i(2), \dots, i(t)$, where $i(j) \notin K(y_i)$ for $j = 1, 2, \dots, t$. Assume the latter holds; note that this is the case for $t = 1$ if i is not a terminal node.

If $i(t) \in I$ then, since $i(t) \notin K(y_i)$, player 1 under $X^*(i)$ chooses the first element not belonging to $K(y_i)$. Thus either play reaches a terminal node giving player 1 a payoff greater than y_i or play is sent to a node $i(t+1) \notin K(y_i)$.

If $i(t) \in J$ then the fact that $i(t) \notin K(y_i)$ implies that, whatever the strategy Y ,

play must either reach a terminal node giving a payoff greater than y_i to player 1 or be sent to a node $i(t+1) \notin K(y_i)$.

Hence at stage $t+1$ we have similar conditions to those at stage t and the process can be repeated. Note that if play never reaches a terminal node the payoff is zero, which is greater than y_i . Hence when $p(i) = 0$ and $n(i) \neq -\infty$ player 1 can ensure himself of a payoff of at least $n(i) - \epsilon$.

Applying similar arguments to the corresponding strategy to $X^*(i)$ for player 2 shows that player 1's payoff can be restricted to at most $p(i) + \epsilon$ if $p(i) \neq 0$, $n(i) + \epsilon$ if $p(i) = 0$ and $n(i)$ is finite and $-1/\epsilon$ when $n(i) = -\infty$. The theorem now follows.

Note. It is not difficult to adapt the proof to show that, if $p(i) > 0$ and $\sup\{x : i \in H(x)\}$ is attained, then player 1 actually has a pure *optimal* stationary strategy in the game starting at node i . Similarly, when $n(i) < 0$ and $\inf\{y : i \in K(y)\}$ is attained, player 2 has a pure *optimal* stationary strategy in the game starting at node i .

4. An example. In this section we present an example of a countably infinite deterministic graphical game expressed in recursive game notation that demonstrates a fundamental difference between finite and infinite deterministic graphical games. In the former, a player possesses a pure stationary strategy that is optimal wherever the game starts whereas, in the latter, this does not necessarily hold.

The following example also shows that, when players are restricted to using Markov strategies, a countably infinite recursive game of perfect information need not have a solution. (A strategy F is said to be Markov if, for all k and all h_k of the form (1), $F^k(h_k)$ depends only on the stage k and the node x_k that play has reached). Strictly speaking, we do not present it in the form of a countably infinite recursive matrix game $\Gamma_1, \Gamma_2, \dots$, but it can clearly be put into this form.

Let Γ be the recursive game given for $n = 1, 2, \dots$ and $i = 1, 2, \dots$ by

$$\begin{aligned} \Gamma_1^{(n)} : (n), \quad \Gamma_{2i}^{(n)} : \left(\begin{array}{c} n-1 \\ \Gamma_{2i-1}^{(n)} \end{array} \right), \\ \Gamma_{2i+1}^{(n)} : (\Gamma_1^{(n)}, \Gamma_2^{(n)}, \dots, \Gamma_{2i}^{(n)}, \Gamma_2^{(n+1)}, \Gamma_4^{(n+1)}, \Gamma_6^{(n+1)}, \dots). \end{aligned}$$

We first use Theorem 3.1 to show that, when play starts in $\Gamma_i^{(n)}$ the value is n . In the deterministic graphical game formulation the nodes $W_i^{(n)}$ are identified with the $\Gamma_i^{(n)}$ in the recursive game formulation, where we also identify $W_1^{(n)}$ with the number entry n in the matrix of $\Gamma_1^{(n)}$. Nodes of the form $W_1^{(n)}$ are seen to be terminal nodes with payoff n . It is easily verified that

$$\begin{aligned} H_0(x) &= \{W_1^{(n)} : n \geq x\}, \\ H_1(x) &= H_0(x) \cup \{W_{2i}^{(n)} : n \geq x+1\} \cup \{W_2^{(n)} : n \geq x\}, \\ H_r(x) &= H_1(x) \cup \bigcup_{j=3}^{r+1} \{W_j^{(n)} : n \geq x\} \cup \bigcup_{j=3}^{2r-1} \{W_j^{(n)} : n \geq x+1\} \quad \text{for } r \geq 2, \end{aligned}$$

so that

$$H(x) = \bigcup_{i \geq 1} \bigcup_{n \geq x} W_i^{(n)}.$$

Hence $p(W_i^{(n)}) = n$ and the value of the game is n .

We assert that Γ does not have a solution in Markov strategies. Let $\epsilon > 0$ and X be a Markov strategy for player 1; we let $x_{2i}^{(n)}(t)$ denote the probability under X of choosing the first row when in $\Gamma_{2i}^{(n)}$ at time t . Let $\delta_1, \delta_2, \dots$ be a decreasing sequence of probabilities such that

$$\delta_1 + 2\delta_2 + 3\delta_3 + \dots < \frac{1}{2}.$$

Let $k_i = \lceil \ln(1 - \epsilon) / \ln(1 - \delta_i) \rceil + 1$ for $i \geq 1$, $t_2 = 1$, and $t_{r+1} = t_r + 2k_r$ for $r \geq 2$. Note that $(1 - \delta_i)^{k_i} < 1 - \epsilon$ for $i \geq 1$ and k_1, k_2, \dots is an increasing sequence.

(i) Suppose there is a positive integer $n \geq 2$ such that $x_{6k_{n+1}+2k_n-2j}^{(n)}(t_n + 2j + 1) \geq \delta_n$ for $j = 0, 1, \dots, k_n - 1$. Consider the Markov strategy Y for player 2 which, for all positive integers r and i , chooses $\Gamma_{2i-1}^{(r)}$ when in $\Gamma_{2i+1}^{(r)}$ at time $t < t_n$ and chooses $\Gamma_{2i}^{(r)}$ when in $\Gamma_{2i+1}^{(r)}$ at time $t \geq t_n$. Under the strategies X and Y , when play starts at $t = 1$ in $\Gamma_{6k_{n+1}+2k_n+2t_n-1}^{(n)}$, play will be in $\Gamma_{6k_{n+1}+2k_n}^{(n)}$ at time $t_n + 1$ and it follows that the expectation is at most

$$(1 - \delta_n)^{k_n} n + \{1 - (1 - \delta_n)^{k_n}\}(n - 1) = n - 1 + (1 - \delta_n)^{k_n} < n - \epsilon.$$

Hence in this case X is not an ϵ -optimal strategy.

(ii) Now suppose (i) does not hold then, for every positive integer $n \geq 2$, we may choose an integer j_n satisfying $0 \leq j_n < k_n$ and $x_{6k_{n+1}+2k_n-2j_n}^{(n)}(t_n + 2j_n + 1) \leq \delta_n$. Consider the Markov strategy Y for player 2 whereby, for all positive integers i and n , he chooses $\Gamma_{6k_{n+2}+2k_{n+1}-2j_{n+1}}^{(n+1)}$ when in $\Gamma_{2i+1}^{(n)}$ at time $t_{n+1} + 2j_{n+1}$ and $\Gamma_{2i-1}^{(n)}$ when in $\Gamma_{2i+1}^{(n)}$ at other times.

Consider play under X and Y when play starts at $t = 1$ in $\Gamma_{6k_2+3}^{(1)}$. At time $t_2 + 2j_2 = 1 + 2j_2$ play is in $\Gamma_{6k_2+3-4j_2}^{(1)}$ so play at time $t_2 + 2j_2 + 1$ is in $\Gamma_{6k_3+2k_2-2j_2}^{(2)}$. Suppose, for $n \geq 2$, at time $t_n + 2j_n + 1$ play is in $\Gamma_{6k_{n+1}+2k_n-2j_n}^{(n)}$.

(a) With probability $x_{6k_{n+1}+2k_n-2j_n}^{(n)}(t_n + 2j_n + 1) \leq \delta_n$ play will terminate giving player 1 a payoff of $n - 1$.

(b) With probability $1 - x_{6k_{n+1}+2k_n-2j_n}^{(n)}(t_n + 2j_n + 1)$ play goes to

$$\Gamma_{6k_{n+1}+2k_n-2j_n-1}^{(n)}$$

at time $t_n + 2j_n + 2 \leq t_n + 2k_n = t_{n+1}$. Hence, for $i = 1, 2, \dots, 2k_n + 2j_{n+1} - 2j_n - 2 = t_{n+1} + 2j_{n+1} - t_n - 2j_n - 2$, at time $t_n + 2j_n + 2 + i$ play is in $\Gamma_{6k_{n+1}+2k_n-2j_n-1-2i}^{(n)}$.

In particular, play is in $\Gamma_{6k_{n+1}-2k_n+2j_n-4j_{n+1}+3}^{(n)}$ at time $t_{n+1} + 2j_{n+1}$; we note that $6k_{n+1} - 2k_n + 2j_n - 4j_{n+1} + 3 \geq 3$. Hence at time $t_{n+1} + 2j_{n+1} + 1$ play is in $\Gamma_{6k_{n+2}+2k_{n+1}-2j_{n+1}}^{(n+1)}$ and the process is repeated. It follows that the expectation is at most

$$\delta_2 + 2\delta_3 + 3\delta_4 + \dots \leq \delta_1 + 2\delta_2 + 3\delta_3 + \dots < \frac{1}{2}$$

and X is not ϵ -optimal when $\epsilon < \frac{1}{2}$.

5. Conclusions. When the restriction that the nodes in a deterministic graphical game be finite is removed, our results show that not only does the corresponding game have a solution but also its solution inherits many of the characteristics of the original game. There is the natural difference that the players may have to be content with ϵ -optimal strategies. The more fundamental change is that a player's ϵ -optimal

strategy may need to depend on the starting node. However once he has this information, the player will be able to employ a pure stationary strategy.

In [11] Washburn briefly considered the possibility of introducing random moves into deterministic graphical games and pointed out that they still have solutions with stationary strategies. In the infinite case such a game need not even have a solution as the following example (given in recursive form) clearly shows when play starts in Γ_1 .

Example. Let $\Gamma = (\Gamma_1, \Gamma_2, \dots)$ be given by

$$\Gamma_1 : ((1/2)\Gamma_2 + (1/2)\Gamma_3), \quad \Gamma_{2n} : \begin{pmatrix} n \\ \Gamma_{2n+2} \end{pmatrix}, \quad \Gamma_{2n+1} : (-n \quad \Gamma_{2n+3}) \text{ for } n \geq 1.$$

REFERENCES

- [1] V. J. BASTON AND F. A. BOSTOCK, *On Washburn's deterministic graphical games*, Differential Games—Developments in Modelling and Computation, R. P. Hämmäläinen and H. K. Ehtamo, eds., Lecture Notes in Control Inform. Sci., Vol. 156, Springer-Verlag, Berlin, 1991, pp. 164–170.
- [2] M. DAVIS, *Infinite games of perfect information*, Advances in Game Theory, M. Dresher, L. S. Shapley, and A. W. Tucker, eds., Ann. Math. Stud., Vol. 52, Princeton University Press, Princeton, NJ, 1964, pp. 85–101.
- [3] A. EHRENFEUCHT AND J. MYCIELSKI, *Positional strategies for mean payoff games*, Internat. J. Game Theory, 8 (1979), pp. 109–113.
- [4] H. EVERETT, *Recursive games*, Contributions to the Theory of Games III, M. Dresher, A. W. Tucker, and P. Wolfe, eds., Ann. Math. Stud., Vol. 39, Princeton University Press, Princeton, NJ, 1957, pp. 47–78.
- [5] D. GALE AND F. M. STEWART, *Infinite games with perfect information*, Contributions to the Theory of Games II, H. W. Kuhn and A. W. Tucker, eds., Ann. Math. Stud., Vol. 28, Princeton University Press, Princeton, NJ, 1953, pp. 245–266.
- [6] J. MYCIELSKI, *Continuous games with perfect information*, Advances in Game Theory, M. Dresher, L. S. Shapley, and A. W. Tucker, eds., Ann. Math. Stud., Vol. 52, Princeton University Press, Princeton, NJ, 1964, pp. 103–112.
- [7] A. S. NOWAK AND T. E. S. RAGHAVAN, *Positive stochastic games and a theorem of Ornstein*, Stochastic Games and Related Topics, T. E. S. Raghavan, T. S. Ferguson, T. Parthasarathy, O. J. Vrieze, Kluwer Academic Publishers, 1991, pp. 127–134.
- [8] J. C. OXToby, *The Banach–Mazur game and Banach category theorem*, Contributions to the Theory of Games III, M. Dresher, A. W. Tucker, and P. Wolfe, eds., Ann. Math. Stud., Vol. 39, Princeton University Press, Princeton, NJ, 1957, pp. 159–163.
- [9] L. C. THOMAS, *Games, Theory and Applications*, Ellis Horwood, Chichester, 1984.
- [10] J. VON NEUMANN AND O. MORGENSTERN, *The Theory of Games and Economic Behaviour*, Princeton University Press, Princeton, NJ, 1944.
- [11] A. WASHBURN, *Deterministic graphical games*, J. Math. Anal. Appl., 153 (1990), pp. 84–96.

AVERAGING IN LAGRANGE AND MINIMAX PROBLEMS OF OPTIMAL CONTROL*

E. N. BARRON†

This paper is dedicated to Avner Friedman on the occasion of his 60th birthday.

Abstract. The perturbed test function method of Evans [*Proc. Royal Soc. Edinburgh*, 111A (1989), pp. 359–375] is applied to two optimal control problems with fast variables. In the classical problem of Lagrange, which has been considered by Chaplais [*SIAM J. Control. Optim.*, 25 (1987), pp. 767–780], a new and simpler proof of convergence to the averaged problem is given. In addition, it is shown that an entirely different limit problem is obtained when the controls have inertia, i.e., are Lipschitz. The optimal control problem with minimax cost and fast variables is studied in section 2. In this problem the cost functional is not the integral of a running cost, which is smooth, but the max in time of a function of the state, the control, and time. The limit problem is derived in this case also.

Key words. optimal control, minimax problem, averaging of fast variables, viscosity solutions, perturbed test functions

AMS subject classifications. 49C20, 49L25, 35B05

Introduction. An important problem in optimal control theory is the modeling of systems that have at least one component that oscillates rapidly. For example, flight systems subject to weather and other disturbances are of this type. The problem considered in this paper is the determination of the limit problem as the oscillations occur infinitely fast. The limit problem is then used as a macroscopic approximation to the rapidly vibrating system. In the first section the classical problem of Lagrange with a fast time variable, say $\varphi(t, t/\varepsilon, x, z)$, is studied. For simplicity we will assume that the dynamics are 1-periodic in the fast variable although this is not necessary—assuming an average exists in the fast variable is enough. The class of controls is, as usual, the class of Lebesgue measurable functions. This problem was first studied by Chaplais [11] (see also Peng [15]) but by completely different methods and with more stringent assumptions in some cases. Chaplais proved that the optimal control limit problem involves the averaged dynamics $\int_0^1 \varphi(t, s, x, z(t, s)) ds$. The important point discovered by Chaplais is that the control z must also involve the fast variable, and an example is given proving that this is indeed necessary. In the dynamic programming approach we take in this paper it becomes very clear why the control must depend on s .

When we are considering problems with fast variables, we must also consider very carefully the class of control functions. The result of Chaplais requires measurable control functions that have the property that they can switch quickly enough (instantaneously) to keep up with the fast variables. It is known, however, that real controls are rarely, if ever, capable of instantaneous changes. Real controls are, in the terminology of Berkovitz, inertial. Consequently, practical controls must be chosen out of the class of Lipschitz-continuous functions. The problem arises as to whether

*Received by the editors March 25, 1991; accepted for publication (in revised form) May 14, 1992. This research was partially supported by Air Force Office of Scientific Research grant AFOSR-86-0202 and National Science Foundation grant DMS- 9102967.

† Department of Mathematical Sciences, Loyola University of Chicago, Chicago, Illinois, 60626.

Lipschitz controls can keep up with the fast variables and thus achieve the same value for the limit problem with measurable controls. The answer given in §1.1 is negative. The limit problem with Lipschitz controls involves the averaged dynamics of the form $\int_0^1 \varphi(t, s, x, z) ds$, where the z control does *not* depend on s .

In §1 we rederive the result of Chaplais (actually only one of his results) using the very powerful and simple perturbed test function method of Evans [13]. This method is intimately connected with another very powerful idea—the idea of viscosity solutions for fully nonlinear partial differential equations [12]. Chaplais also uses viscosity solution theory to prove convergence of the value functions but he resorts to second-order approximations, which are not, in fact, necessary.

The perturbed test function method of Evans is connected with the older known method of expansion in ε , i.e., looking at $V^\varepsilon(t, x) = V^0(t, x) + \varepsilon V^1(\frac{t}{\varepsilon}, x) + \dots$. This looks at perturbing the *solution*, which is usually not smooth enough. Evans instead looks at perturbing the (smooth) test function in the definition of viscosity solution in an appropriate way. Then we can apply the usual tricks associated with the older method. Therein lies the utility of the method. His method is very successfully applied in [13] for second-order problems. However, for first-order problems, the equation for the perturbation does not generally have a smooth enough solution to generate a test function that can be used in the viscosity definition. This is, in fact, the case for homogenization (see §1.3), but is not true for a fast time variable. In that case, a smooth perturbation can be constructed (see (1.13)).

In §2 we study the optimal control problem in which the cost is of the form $P(\zeta) = \sup_{t \leq \tau \leq T} h(\tau, \xi(\tau), \zeta(\tau))$. When the ζ controls are chosen to minimize the cost, this is called the minimax problem.

The minimax control problem was first considered from the dynamic programming point of view in [6]. (It later came to our attention that Aronsson in [1] had considered the minimax calculus of variations problem from the point of view of the Euler equation). After seeing [1] and [6], the reader will understand why this criterion is not widely used, even though it is usually more realistic than an integral running cost. In fact, for some problems the minimax cost is the only realistic formulation. A recent example of this appears in [10], in which the problem of landing an aircraft in the presence of windshear is studied. For minimax problems, the necessary conditions are much more difficult to solve in virtually all cases. Also, the Bellman equation requires a minimization over a set that involves the solution function (see (2.1) and (2.5)). Nevertheless, the results for this problem are not beyond the scope of numerical methods.

The question we pose in §2 is the determination of the limit of the rapidly oscillating minimax optimal control problem. The limit problem is discovered using the usual L^p approximations to the L^∞ norm. Evans' technique is then used to actually prove the convergence and is presented in Theorem 2.2. The cost function h in this problem is not allowed to depend on the fast variable because it is not clear that such dependence leads to convergence. In fact, looking at the Bellman equation as a nonlinear variational inequality (see (2.1a)), the lower obstacle for the ε -value function, assuming fast variable dependence, is given by $\beta(t, t/\varepsilon, x) \equiv \min_{z \in Z} h(t, t/\varepsilon, x, z)$. The function β does converge in some sense, for example, weakly to the average, or epi-convergence [2] to an infimum, or hypo-convergence to a supremum. None of these seems to be strong enough to give convergence of the value functions. This is an open problem.

When oscillations occur in the spatial parameters of a system the limit equations are known as the homogenized version (see Bensoussan, Lions, and Papanicolaou [9] for the basic results concerning homogenization). Lions, Papanicolaou, and Varadhan in [14] study the asymptotic problems for first-order Hamilton–Jacobi equations of the form $H(x, x/\varepsilon, u, Du) = 0$. This problem is much more difficult than the problem considered in this paper. While a limit problem is proved to exist, the characterization of the limit Hamiltonian is known only in special cases. To show where the difficulties with the perturbed test function method arise, we will consider in §1.2 a simple one-dimensional control problem. This example also establishes that the controls for the homogenized problem must depend on the fast spatial parameter, just as they depend on the fast time parameter for the problems in this paper.

1. The Lagrange optimal control problem. In this section we formulate precisely the standard free endpoint problem of Lagrange in optimal control. The dynamics are

$$(1.1) \quad \frac{d\xi}{d\tau} = f\left(\tau, \frac{\tau}{\varepsilon}, \xi(\tau), \zeta(\tau)\right) \quad \text{if } t < \tau \leq T,$$

$$(1.2) \quad \xi(t) = x \in R^n.$$

The control functions ζ are chosen from the class of functions

$$\mathcal{Z}[t, T] = \{\zeta : [t, T] \rightarrow Z \mid \zeta(\cdot) \text{ is Lebesgue measurable}\},$$

where Z is a compact subset of some R^p , $p \geq 1$. The objective will be to minimize on $\mathcal{Z}[t, T]$ the following payoff:

$$(1.3) \quad P_{t,x}(\zeta) = g(\xi(T)) + \int_t^T h\left(r, \frac{r}{\varepsilon}, \xi(r), \zeta(r)\right) dr.$$

We define the value function $V^\varepsilon : [0, T] \times R^n \rightarrow R^1$ as follows:

$$V^\varepsilon(t, x) = \inf_{\zeta \in \mathcal{Z}[t, T]} P_{t,x}(\zeta).$$

The following assumptions regarding the given functions f , g , and h will be used throughout this paper.

(A). For $\varphi = f, h, \varphi : [0, T] \times R^1 \times R^n \times Z \rightarrow R^1$ is uniformly continuous in all arguments. $\varphi(\cdot, s, \cdot, \cdot)$ is periodic with period 1. There is a generic constant $K > 0$ such that

$$|f(t, s, x, z)| \leq K(1 + |x|), \quad |h(t, s, x, z)| \leq K,$$

for $(t, s, x, z) \in [0, T] \times R^1 \times R^n \times Z$ and

$$|\varphi(t, s, x, z) - \varphi(t', s, x', z)| \leq K(|x - x'| + |t - t'|).$$

Further, $|g(x)| \leq K(1 + |x|)$, and g is also uniformly Lipschitz continuous.

Assumption (A) is sufficient to guarantee that for each control $\zeta \in \mathcal{Z}[t, T]$ there will be a unique trajectory $\xi_\varepsilon = \xi(\cdot)$ on the interval $[t, T]$ with $\xi(t) = x$. The hypotheses can be weakened to allow h to have quadratic growth in x for the Lagrange problem.

To simplify notation, let Ω denote the set $(0, T) \times R^n$.

For the convenience of the reader we recall the precise definition of viscosity solution. In general, both the solution and the Hamiltonian may be discontinuous. The notation is that for any function f , f^* is the upper semicontinuous envelope and f_* is the lower semicontinuous envelope: $f^*(x) = \limsup_{y \rightarrow x} f(y)$, and $f_*(x) = \liminf_{y \rightarrow x} f(y)$.

DEFINITION 1.1. A function $u : \bar{\Omega} \rightarrow R^1$ is a viscosity solution of

$$(1.4) \quad u_t + F(t, x, u, D_x u) = 0$$

if the following two conditions are satisfied:

(1) u is a viscosity subsolution, i.e., for any function $\varphi \in C^1(\Omega)$ for which $u^* - \varphi$ achieves a maximum at the point $(t_0, x_0) \in \Omega$ we have that

$$(1.5) \quad \varphi_t(t_0, x_0) + F^*(t_0, x_0, u^*(t_0, x_0), D_x \varphi(t_0, x_0)) \geq 0 \quad \text{at } (t_0, x_0).$$

(2) u is a viscosity supersolution, i.e., for any function $\varphi \in C^1(\Omega)$ for which $u_* - \varphi$ achieves a minimum at the point $(t_0, x_0) \in \Omega$ we have that

$$(1.6) \quad \varphi_t(t_0, x_0) + F_*(t_0, x_0, u_*(t_0, x_0), D_x \varphi(t_0, x_0)) \leq 0 \quad \text{at } (t_0, x_0).$$

As was established in [12], we may always arrange to have unique strict extrema and $0 < t_0 < T$. Furthermore, it is not necessary to have $\varphi \in C^1$. In fact, differentiability of φ at the point of contact with u^* or u_* is sufficient.

If the Hamiltonian $F(t, x, r, p)$ is continuous in all variables and concave in p then it was established in [8] that a continuous function u is a viscosity solution of (1.4) if and only if the condition " $u - \varphi$ achieves a minimum at (t_0, x_0) " implies that

$$(1.6a) \quad \varphi_t(t_0, x_0) + F(t_0, x_0, u(t_0, x_0), D_x \varphi(t_0, x_0)) = 0 \quad \text{at } (t_0, x_0).$$

That is, the fact that u is a supersolution in this stronger sense is sufficient.

In several places we will use a lemma of Barles and Perthame [3], which we will reproduce here for the convenience of the reader. It will be stated only for subsolutions but is also true for supersolutions, suitably modified.

LEMMA A.3 (see [3]). Let $u(t, x) = \limsup_{(n, \sigma, y) \rightarrow (\infty, t, x)} u_n(\sigma, y)$, where, for all n , u_n is a subsolution of $u_t + F_n(t, x, u, D_x u) = 0$. Assume that u_n and u are locally bounded. Let $\varphi \in C^1(\Omega)$ and (t_0, x_0) the unique maximum of $u - \varphi$ in $B_\varrho(t_0, x_0)$ for some $\varrho > 0$. For each n , let $(t_n, x_n) \in \overline{B_\varrho(t_0, x_0)}$ be a maximum of $u_n - \varphi$. Then $(t_n, x_n) \rightarrow (t_0, x_0)$ and $u_n(t_n, x_n) \rightarrow u(t_0, x_0)$ as $n \rightarrow \infty$.

As pointed out by the reviewer, certain modifications of the test function φ in Lemma A.3 may be necessary so as to keep $(t_n, x_n) \in \overline{B_\varrho(t_0, x_0)}$. The details are to be found in [3].

The next lemma gives the dynamic programming result for the fast problem.

LEMMA 1.2. The value function V^ε is the unique continuous viscosity solution of the problem

$$(1.7) \quad \frac{\partial V}{\partial t} + \min_{z \in Z} \left\{ D_x V \cdot f \left(t, \frac{t}{\varepsilon}, x, z \right) + h \left(t, \frac{t}{\varepsilon}, x, z \right) \right\} = 0 \quad \text{on } \Omega,$$

with terminal condition

$$(1.8) \quad V^\varepsilon(T, x) = g(x) \quad \text{on } R^n.$$

Further, there exists a Lipschitz continuous function $W : \bar{\Omega} \rightarrow R^1$ such that (at least) on a subsequence of $\{\varepsilon\}$ we have

$$(1.9) \quad \lim_{\varepsilon \rightarrow 0} V^\varepsilon(t, x) = W(t, x).$$

Define the Hamiltonian $H : [0, T] \times R^1 \times R^n \times R^n \rightarrow R^1$ by

$$(1.10) \quad H(t, s, x, p) = \min_{z \in Z} \{p \cdot f(t, s, x, z) + h(t, s, x, z)\}.$$

Proof. The fact that V^ε is the unique continuous viscosity solution of the Bellman equation (1.7) is well known (e.g., [7]). Now we secure a convergent subsequence by showing that V^ε is uniformly Lipschitz continuous. Since this is rather standard we will only sketch the proof, needing to make sure that the estimates are uniform in ε .

Let $x, x' \in R^n$ and fix $0 \leq t < T$ and $\zeta \in \mathcal{Z}[t, T]$. Denote the solution of (1.1) with initial condition x or x' by ξ_x and $\xi_{x'}$. Then, using condition (A),

$$|\xi_x(\tau) - \xi_{x'}(\tau)| \leq |x - x'| + \int_t^\tau K |\xi_x(s) - \xi_{x'}(s)| ds.$$

Gronwall's inequality then gives

$$\sup_{t \leq \tau \leq T} |\xi_x(\tau) - \xi_{x'}(\tau)| \leq K|x - x'|.$$

This immediately gives uniform Lipschitz continuity of V^ε in x , and since g is uniformly Lipschitz, this is true on $[0, T]$.

Let $0 \leq t_1 < t_2 \leq T$ and fix $x \in R^n$ and $\zeta \in \mathcal{Z}[0, T]$. Denote the solution of (1.1) with initial condition $\xi_i(t_i) = x, i = 1, 2$. Then, since f is uniformly Lipschitz in x , we easily see that $|\xi_1(t_2) - x| \leq K(t_2 - t_1)$, with K independent of ε . Then

$$\begin{aligned} |\xi_1(\tau) - \xi_2(\tau)| &\leq |\xi_1(t_2) - x| + \int_{t_2}^\tau K |\xi_1(s) - \xi_2(s)| ds \\ &\leq K(t_2 - t_1) + \int_{t_2}^\tau K |\xi_1(s) - \xi_2(s)| ds. \end{aligned}$$

Another application of Gronwall's inequality yields the result that the trajectories cannot separate by more than $K(t_2 - t_1)$ on $[t_2, T]$. Consequently, this is true on $[t_1, T]$. Using now the boundedness and Lipschitz continuity of h and g , Lipschitz continuity of V^ε in t , uniformly in ε is immediate.

We see that $\|V^\varepsilon\|$ and $\|DV^\varepsilon\| \leq K$ in the uniform norm, uniformly in ε . Therefore, a subsequence converges uniformly to a Lipschitz function W . \square

Remarks. 1. Chaplais obtains a convergent subsequence in the following way. Let V_α^ε denote the (classical) solution of

$$\frac{\partial V_\alpha^\varepsilon}{\partial t} + \alpha \Delta V_\alpha^\varepsilon + H\left(t, \frac{t}{\varepsilon}, x, D_x V_\alpha^\varepsilon\right) = 0, \quad (t, x) \in \Omega,$$

with $V_\alpha^\varepsilon(T, x) = g(x)$. It is not hard to show that V_α^ε and its first derivatives are uniformly bounded in an appropriate norm. Since V^ε satisfies the property that

$$\|V^\varepsilon - V_\alpha^\varepsilon\|_{L^\infty(\bar{\Omega})} \leq K\alpha^{1/2},$$

we see that $\{V^\varepsilon\}_\varepsilon$ has a convergent subsequence, to W say, and W is a continuous function. See Chaplais [11, p. 777] for complete details.

2. Observe that in (1.10) any minimizer in Z will depend on t, x, p , and $s \in [0, 1]$. Consequently, it is clear that there must be s dependence of the controls for the limit problem.

Thus is secured a convergent sequence to a nice function W . To describe the equation satisfied by W , introduce the averaged dynamics

$$(1.11a) \quad \bar{f}(t, x, z(\cdot)) = \int_0^1 f(t, s, x, z(s)) \, ds,$$

$$(1.11b) \quad \bar{h}(t, x, z(\cdot)) = \int_0^1 h(t, s, x, z(s)) \, ds,$$

where

$$z(\cdot) \in \mathcal{A} \equiv \{z : [0, 1] \rightarrow Z \mid z \text{ is Lebesgue measurable}\}.$$

Define the averaged Hamiltonian

$$(1.12) \quad \bar{H}(t, x, p) = \min_{z(\cdot) \in \mathcal{A}} \{p \cdot \bar{f}(t, x, z(\cdot)) + \bar{h}(t, x, z(\cdot))\}.$$

The first goal of this section is to provide a new proof of the following result of Chaplais [11] by using the perturbed test function method of Evans [13].

THEOREM 1.3. *The function W is the unique continuous viscosity solution of*

$$(HJB) \quad \frac{\partial W}{\partial t} + \min_{z(\cdot) \in \mathcal{A}} \{D_x W \cdot \bar{f}(t, x, z(\cdot)) + \bar{h}(t, x, z(\cdot))\} = 0, \quad (t, x) \in [0, T] \times R^n,$$

with $W(T, x) = g(x)$.

Remark. Chaplais proves the interesting and useful result that if the limit problem has an optimal control, say $\zeta(t, s)$, then $\zeta(t, t/\varepsilon)$ is near optimal for the ε -oscillation problem.

Proof. This can be done by simply showing (1.6a), but we will go through both (1.5) and (1.6). First, we need the lemma.

LEMMA 1.3a. *For $(t, x, p) \in [0, T] \times R^n \times R^n$, we have that*

$$\int_0^1 H(t, s, x, p) \, ds = \bar{H}(t, x, p).$$

Proof. Fix (t, x, p) . By Fillipov's lemma, since Z is compact, there exists a measurable function $\zeta : [0, 1] \rightarrow Z$ such that

$$H(t, s, x, p) = p \cdot f(t, s, x, \zeta(s)) + h(t, s, x, \zeta(s)).$$

Integrating both sides we get

$$\begin{aligned} \int_0^1 H(t, s, x, p) \, ds &= \int_0^1 p \cdot f(t, s, x, \zeta(s)) + h(t, s, x, \zeta(s)) \, ds \\ &\geq \min_{z(\cdot) \in \mathcal{A}} \{p \cdot \bar{f}(t, x, z(\cdot)) + \bar{h}(t, x, z(\cdot))\} = \bar{H}(t, x, p). \end{aligned}$$

On the other hand, for fixed (t, x, p) , for any $\delta > 0$ there exists $z_\delta \in \mathcal{A}$ such that

$$\begin{aligned}\overline{H}(t, x, p) &\geq p \cdot \bar{f}(t, x, z_\delta(\cdot)) + \bar{h}(t, x, z_\delta(\cdot)) - \delta \\ &= \int_0^1 p \cdot f(t, s, x, z_\delta(s)) + h(t, s, x, z_\delta(s)) \, ds - \delta \\ &\geq \int_0^1 H(t, s, x, p) \, ds - \delta. \quad \square\end{aligned}$$

Now, suppose that $W - \varphi$ has a strict maximum at a point $(t_0, x_0) \in \Omega$, where $\varphi \in C^2(\Omega)$. Consider the function $\gamma : R^1 \rightarrow R^1$ given as the 1-periodic solution of

$$\begin{aligned}(1.13) \quad \frac{d\gamma}{ds} &= \overline{H}(t_0, x_0, D_x \varphi(t_0, x_0)) - H(t_0, s, x_0, D_x \varphi(t_0, x_0)) \\ &= \min_{z(\cdot) \in \mathcal{A}} \{D_x \varphi(t_0, x_0) \cdot \bar{f}(t_0, x_0, z(\cdot)) + \bar{h}(t_0, x_0, z(\cdot))\} \\ &\quad - \min_{z \in Z} \{D_x \varphi(t_0, x_0) \cdot f(t_0, s, x_0, z) + h(t_0, s, x_0, z)\}\end{aligned}$$

if $0 \leq s \leq 1$. A periodic solution is guaranteed to exist because of Lemma 1.3a. Note that the right-hand side of (1.14) is continuous in $s \in [0, 1]$ and, since f and h are periodic in s , $\gamma'(0) = \gamma'(1)$. Therefore, $\gamma \in C^1(R^1)$.

Since $V^\varepsilon \rightarrow W$ uniformly, there exists a sequence $(t_\varepsilon, x_\varepsilon)$ such that $V^\varepsilon(t, x) - (\varphi(t, x) + \varepsilon\gamma(t/\varepsilon))$ achieves a maximum at $(t_\varepsilon, x_\varepsilon)$ and $(t_\varepsilon, x_\varepsilon) \rightarrow (t_0, x_0)$ as $\varepsilon \rightarrow 0$.

Since V^ε is a viscosity solution of (1.7), we apply the definition for subsolution to get that

$$(1.14) \quad \varphi_t(t_\varepsilon, x_\varepsilon) + \frac{d\gamma(\frac{t_\varepsilon}{\varepsilon})}{dt} + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \geq 0,$$

with H given in (1.10). Using the definition (1.13) for γ evaluated at $s = t_\varepsilon/\varepsilon$, we obtain from (1.14) and the continuity of H ,

$$\begin{aligned}0 &\leq \varphi_t(t_\varepsilon, x_\varepsilon) + \frac{d\gamma(\frac{t_\varepsilon}{\varepsilon})}{dt} + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \\ &= \varphi_t(t_\varepsilon, x_\varepsilon) + \overline{H}(t_0, x_0, D_x \varphi(t_0, x_0)) - \\ &\quad H\left(t_0, \frac{t_\varepsilon}{\varepsilon}, x_0, D_x \varphi(t_0, x_0)\right) + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \\ &\leq \varphi_t(t_0, x_0) + \overline{H}(t_0, x_0, D_x \varphi(t_0, x_0)) + o_\varepsilon(1)\end{aligned}$$

as $\varepsilon \rightarrow 0$, since $(t_\varepsilon, x_\varepsilon) \rightarrow (t_0, x_0)$ and φ is smooth. We conclude that

$$(1.15) \quad \varphi_t(t_0, x_0) + \min_{z(\cdot) \in \mathcal{A}} \{D_x \varphi(t_0, x_0) \cdot \bar{f}(t_0, x_0, z(\cdot)) + \bar{h}(t_0, x_0, z(\cdot))\} \geq 0$$

and so W is a subsolution of (HJB).

Next we prove in a similar way that W is a supersolution.

Suppose that $W - \varphi$ has a strict minimum at (t_0, x_0) . Let $\gamma : R^1 \rightarrow R^1$ be the 1-periodic solution of (1.13). $V^\varepsilon - (\varphi(t, x) + \varepsilon\gamma(t/\varepsilon))$ achieves a minimum at

$(t_\varepsilon, x_\varepsilon) \rightarrow (t_0, x_0)$, and since V^ε is a supersolution of (1.7) we have from (1.13) with $s = t_\varepsilon/\varepsilon$

$$\begin{aligned} 0 &\geq \varphi_t(t_\varepsilon, x_\varepsilon) + \frac{d\gamma(\frac{t_\varepsilon}{\varepsilon})}{dt} + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, D_x\varphi(t_\varepsilon, x_\varepsilon)\right) \\ &= \varphi_t(t_\varepsilon, x_\varepsilon) + \bar{H}(t_0, x_0, D_x\varphi(t_0, x_0)) - \\ &\quad H\left(t_0, \frac{t_\varepsilon}{\varepsilon}, x_0, D_x\varphi(t_0, x_0)\right) + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, D_x\varphi(t_\varepsilon, x_\varepsilon)\right) \\ &\geq \varphi_t(t_0, x_0) + \bar{H}(t_0, x_0, D_x\varphi(t_0, x_0)) - o_\varepsilon(1) \end{aligned}$$

as $\varepsilon \rightarrow 0$. Consequently,

$$(1.16) \quad \varphi_t(t_0, x_0) + \min_{z(\cdot) \in \mathcal{A}} \{D_x\varphi(t_0, x_0) \cdot \bar{f}(t_0, x_0, z(\cdot)) + \bar{h}(t_0, x_0, z(\cdot))\} \leq 0.$$

This says that W is a supersolution of (HJB) as well.

Since $V^\varepsilon(T, x) = g(x)$ for all $\varepsilon > 0$, $W(T, x) = g(x)$, $x \in R^n$. Combining this with (1.15) and (1.16) we have established the theorem. \square

Remarks. As a general remark, the critical property of the perturbed test function γ is that it is bounded, smooth, and periodic.

Another proof of Lemma 1.3a using control theory is the following. Consider the optimal control problem: Minimize over controls $z(\cdot) \in \mathcal{A}$ the cost functional $p \cdot \int_0^1 f(t, r, x, z(r)) + h(t, r, x, z(r)) \, dr$, where $t \in [0, T]$, $p, x \in R^n$ are fixed. Let $y \in [0, 1]$ and define the value function

$$(1.17) \quad U(y) = \min_{z(\cdot) \in \mathcal{A}} p \cdot \int_y^1 f(t, r, x, z(r)) + h(t, r, x, z(r)) \, dr.$$

Then U is the unique viscosity solution of

$$(1.18) \quad \frac{dU}{dy} + \min_{z \in Z} \{p \cdot f(t, y, x, z) + h(t, y, x, z)\} = 0, \quad U(1) = 0.$$

However, the $(C^1(0, 1))$ solution of this ordinary differential equation is obviously $\int_y^1 \min_{z \in Z} \{p \cdot f(t, r, x, z) + h(t, r, x, z)\} \, dr$. Uniqueness then says that

$$\min_{z(\cdot) \in \mathcal{A}} p \cdot \int_y^1 f(t, r, x, z(r)) + h(t, r, x, z(r)) \, dr = \int_y^1 \min_{z \in Z} \{p \cdot f(t, r, x, z) + h(t, r, x, z)\} \, dr,$$

for all $y \in [0, 1]$.

Finally, it is straightforward to verify that the simple integral problem in (1.17) always has an optimal control and it can be found in feedback form from (1.18). Consequently, the averaged Hamiltonian \bar{H} is always well defined, i.e., the minimum is attained.

The following corollary is a consequence of the fact that W is the unique viscosity solution of (HJB) and the fact that the value function for the optimal control problem (1.19)-(1.20) also is a viscosity solution of (HJB).

COROLLARY 1.4. *The function W in Theorem 1.3 is the value function for the optimal control problem with dynamics*

$$(1.19) \quad \frac{d\xi}{d\tau} = \int_0^1 f(\tau, s, \xi(\tau), \zeta(\tau, s)) ds, t < \tau \leq T, \quad \xi(t) = x \in R^n,$$

and cost

$$(1.20) \quad \bar{P}_{t,x}(\zeta) = g(\xi(T)) + \int_t^T \int_0^1 h(r, s, \xi(r), \zeta(r, s)) ds dr,$$

where $\zeta(\tau, \cdot) \in \mathcal{Z}[t, T], \zeta(\cdot, s) \in \mathcal{A}$. We have

$$W(t, x) = \inf_{\zeta(\cdot, \cdot)} \bar{P}_{t,x}(\zeta(\cdot, \cdot)).$$

Consequently, we have established that the limit of the fast optimal control problem is the control problem associated with (1.19), (1.20) by using the uniqueness of viscosity solutions. This avoids the technical complexities of a direct proof.

Remark. The result of this section can be extended to differential games. For example, the fast upper value function $V^{\varepsilon+}$ (assuming that $h = 0$) is the viscosity solution of the Isaacs equation

$$\frac{\partial V^{\varepsilon+}}{\partial t} + \min_{z \in Z} \max_{y \in Y} \left(D_x V^{\varepsilon+} f\left(t, \frac{t}{\varepsilon}, x, y, z\right) \right) = 0.$$

Using the perturbed test function

$$\frac{d\gamma}{ds} = \min_{z(\cdot)} \max_{y(\cdot)} (\varphi_x \bar{f}(t, x, y(\cdot), z(\cdot))) - \min_{z \in Z} \max_{y \in Y} (\varphi_x f(t, s, x, y, z))$$

we prove that $V^{\varepsilon+} \rightarrow W$ and W satisfies

$$W_t + \min_{z(\cdot)} \max_{y(\cdot)} (W_x \bar{f}(t, x, y(\cdot), z(\cdot))) = 0.$$

A similar statement holds for the lower value. In the differential game case, both players z , the minimizer, and y , the maximizer, must depend on the fast variable, in general.

Remark. If we are concerned with the stochastic optimal control problem

$$\frac{d\xi}{d\tau} = f\left(\tau, \frac{\tau}{\varepsilon}, \xi(\tau), \zeta(\tau)\right) d\tau + \sigma\left(\tau, \frac{\tau}{\varepsilon}, \xi(\tau), \zeta(\tau)\right) dw(\tau), \quad \xi(t) = x \in R^1,$$

$$V^\varepsilon(t, x) = \inf_{\zeta} E_{t,x}[g(\xi(T))],$$

where $w(\cdot)$ is Brownian motion, under suitable assumptions on f , g , σ , it is not difficult to show that $V^\varepsilon \rightarrow W$ on a subsequence, and W is Lipschitz continuous. Then, by the same method of proof of theorem 1.3 using the perturbation γ which solves

$$\begin{aligned} \frac{d\gamma}{ds} = \min_{z(\cdot)} \left[\frac{1}{2} \int_0^1 \sigma^2(t_0, r, x_0, z(r)) \varphi_{xx}(t_0, x_0) dr + \int_0^1 f(t_0, r, x_0, z(r)) \varphi_x(t_0, x_0) dr \right] \\ - \min_{z \in Z} \left[\frac{1}{2} \sigma^2(t_0, s, x_0, z) \varphi_{xx}(t_0, x_0) + f(t_0, s, x_0, z) \varphi_x(t_0, x_0) \right], \\ 0 \leq s \leq 1, \end{aligned}$$

with γ 1-periodic, it is easily proved that W solves, in the viscosity sense,

$$W_t + \min_{z(\cdot)} \left[\frac{1}{2} W_{xx}(t, x) \int_0^1 \sigma^2(t, s, x, z(s)) ds + W_x(t, x) \int_0^1 f(t, s, x, z(s)) ds \right] = 0,$$

and $W(T, x) = g(x)$. This is also true in R^n . Note that σ^2 is the averaged diffusion, not σ . Also note that the perturbation γ is continuously differentiable and that this is sufficient for a test function against W in time.

1.1 What happens if we use Lipschitz Controls?. In virtually any real problem, measurable controls cannot actually be implemented since most systems are inertial. That is, instantaneous changes in the control and the state are impossible. In effect, this says that practical problems involve the use of Lipschitz-continuous controls, not measurable controls. If that is the case, then it is intuitive that the Lipschitz control cannot keep up with rapidly oscillating dynamics. The question then arises as to exactly what can be achieved. The answer is contained in Theorem 1.5.¹

For simplicity, assume that the control set $Z = [0, 1]^p$ and extend the dynamics f and h so that they are 1-periodic in the z variable. We need to do something like this because of the following.

When the control $\zeta : [0, T] \rightarrow Z$ is Lipschitz, say with Lipschitz constant $L > 0$, we can consider the derivative of ζ to be the real control. Thus we have the dynamics

$$\begin{aligned} \frac{d\xi}{d\tau} &= f\left(\tau, \frac{\tau}{\varepsilon}, \xi(\tau), \zeta(\tau)\right), \\ \frac{d\zeta}{d\tau} &= \beta(\tau) \quad \text{if } t < \tau \leq T, \\ \xi(t) &= x \in R^n, \quad \zeta(t) = z \in Z. \end{aligned}$$

The control functions β are chosen from the class of functions

$$B_L[t, T] = \{\beta : [t, T] \rightarrow [-L, L]^p \mid \beta(\cdot) \text{ is Lebesgue measurable}\}.$$

First, note that we must impose an initial condition on ζ . When the controls ζ are merely measurable, this is irrelevant because the control can instantaneously jump to any desired point. However, when they are Lipschitz this is impossible. On the other hand, when the Lipschitz constant becomes arbitrarily large, it seems reasonable, and is indeed true, that the dependence on the initial condition $z \in Z$ goes away.

Second, in this formulation of the Lipschitz control problem, ζ becomes a state variable. This is, therefore, a state constrained problem due to the fact that we require $\zeta \in Z[t, T]$. By making the dynamics periodic in z , however, we can remove the state constraint because the action of any control ζ that leaves Z can be obtained via a control that does not exit via reflection at the boundaries of $Z = [0, 1]^p$. In this way, if we start in Z we will stay in Z .

Even though periodic extension may destroy continuity of the dynamics, this does not cause a problem. On the other hand, if we do not extend the dynamics, we would need to impose conditions on the value function at the boundary of Z . These conditions are, of course, unknown in general.

Complete details and further results regarding the Lipschitz control problem can be found in [4], [5], [7].

Now we turn to our interest in Lipschitz controls in the context of this paper.

¹This is the outcome of a fruitful conversation with Bob Jensen.

THEOREM 1.5. *For each $L > 0$, $\varepsilon > 0$, let $V^{\varepsilon,L} = V^{\varepsilon,L} : [0, T] \times R^n \times Z \rightarrow R^1$ be the value function for the optimal control problem (1.1)–(1.4) when the controls ζ must be chosen from*

$$\mathcal{Z}_z^L[t, T] = \{\zeta \in \mathcal{Z}[t, T] \mid |\zeta(\tau) - \zeta(\tau')| \leq L|t - t'|, \forall \tau, \tau' \in [t, T], \zeta(t) = z\}.$$

Then

- (1) $\lim_{\varepsilon \rightarrow 0} \lim_{L \rightarrow \infty} V^{\varepsilon,L}(t, x, z) = W(t, x)$ where W is the solution of (HJB).
- (2) $\lim_{L \rightarrow \infty} \lim_{\varepsilon \rightarrow 0} V^{\varepsilon,L}(t, x, z) = U(t, x)$ where U is the value function for the optimal control problem with dynamics

$$\frac{d\xi}{d\tau} = \int_0^1 f(\tau, s, \xi(\tau), \zeta(\tau)) ds, \quad t < \tau \leq T, \quad \xi(t) = x \in R^n,$$

and cost

$$P_{t,x}(\zeta) = g(\xi(T)) + \int_t^T \int_0^1 h(r, s, \xi(r), \zeta(r)) dr, \quad \zeta \in \mathcal{Z}[t, T].$$

U is the viscosity solution of

$$(HJB^*) \quad U_t + \min_{z \in Z} \{D_x U \cdot \int_0^1 f(t, s, x, z) ds + \int_0^1 h(t, s, x, z) ds\} = 0,$$

and $U(T, x) = g(x)$.

Note that the controls in the limit problem yielding U do *not* depend on the fast variable s .

Proof. A special case of [7, Thm. 4.1] says that for each fixed $L > 0$, $\varepsilon > 0$, $V^{\varepsilon,L}$ is the unique continuous viscosity solution of

$$(1.21) \quad \begin{aligned} V_t^{\varepsilon,L} + D_x V^{\varepsilon,L} \cdot f\left(t, \frac{t}{\varepsilon}, x, z\right) + h\left(t, \frac{t}{\varepsilon}, x, z\right) - L|D_z V^{\varepsilon,L}| &= 0, \quad (t, x) \in \Omega, z \in Z \\ V^{\varepsilon,L}(T, x, z) &= g(x). \end{aligned}$$

Now, to prove (1) we use the fact ([7]) that for fixed ε , as $L \rightarrow \infty$, $V^{\varepsilon,L}(t, x, z) \rightarrow V^\varepsilon(t, x)$, where V^ε is the unique viscosity solution of (1.7) and is the same as the value function of Lemma 1.2. Then, as $\varepsilon \rightarrow 0$, $V^\varepsilon \rightarrow W$, according to Theorem 1.3. Thus (1) is proved. This says that if we first let the Lipschitz constant be arbitrarily large then the controls can keep up with the oscillations. This is the measurable control case. This will be in sharp distinction to part (2) in which we first take the fast oscillations for a fixed Lipschitz constant and then let the Lipschitz constant become arbitrarily large.

To prove (2), write (1.21) as

$$V_t^{\varepsilon,L} + \min_{|a| \leq L} \left\{ D_x V^{\varepsilon,L} \cdot f\left(t, \frac{t}{\varepsilon}, x, z\right) + h\left(t, \frac{t}{\varepsilon}, x, z\right) + a D_z V^{\varepsilon,L} \right\} = 0.$$

Fix $L > 0$ and let $\varepsilon \rightarrow 0$. Using Theorem 1.3, we obtain $V^{\varepsilon,L}(t, x, z) \rightarrow U^L(t, x, z)$, where U^L satisfies

$$U_t^L + \min_{|a(\cdot)| \leq L} \left\{ D_x U^L \cdot \int_0^1 f(t, s, x, z) ds + \int_0^1 h(t, s, x, z) ds + D_z U^L \cdot \int_0^1 a(s) ds \right\} = 0.$$

The min is taken over functions $|a(\cdot)| \leq L$. However, this is the same as

$$U_t^L + D_x U^L \cdot \int_0^1 f(t, s, x, z) ds + \int_0^1 h(t, s, x, z) ds - L|D_z U^L| = 0.$$

Now let $L \rightarrow \infty$. Using again [7, Thm. 4.2], we get $U^L(t, x, z) \rightarrow U(t, x)$, where U is the unique viscosity solution of (HJB*). Since the value function in the statement of (2) is known to be the viscosity solution of this equation, U must be the value function. This completes the proof. \square

COROLLARY 1.6. $W(t, x) \leq U(t, x)$ on $[0, T] \times R^n$.

Proof. The corollary follows immediately from the fact that

$$\begin{aligned} \min_{z(\cdot) \in \mathcal{A}} \left\{ p \cdot \int_0^1 f(t, s, x, z(s)) ds + \int_0^1 h(t, s, x, z(s)) ds \right\} \\ \leq \min_{z \in Z} \left\{ p \cdot \int_0^1 f(t, s, x, z) ds + \int_0^1 h(t, s, x, z) ds \right\}. \quad \square \end{aligned}$$

Remark. The result is that in practical problems we should not expect to achieve the averaged value W . The best we should expect is the value U .

1.2. A Homogenized problem in one dimension. In this section we will discuss the homogenized problem for a fast variable in the state, not in time. Specifically, the fast control problem is

$$\begin{aligned} \frac{d\xi}{d\tau} &= f\left(\tau, \frac{\xi}{\varepsilon}, \xi(\tau), \zeta(\tau)\right) \quad \text{if } t < \tau \leq T, \\ \xi(t) &= x \in R^1. \end{aligned}$$

We will assume that (A) holds and that the real valued function $f(\cdot, y, \cdot, \cdot)$ is uniformly Lipschitz and 1-periodic. Also assume that $\min_{z \in Z} f(t, y, x, z) \geq \alpha > 0$. The control functions for the fast problem ζ are chosen from the class $Z[t, T]$. For simplicity, the objective will be to minimize on $Z[t, T]$ the Mayer payoff $P_{t,x}(\zeta) = g(\xi(T))$. The fast value function is $V^\varepsilon(t, x) = \inf_{\zeta \in Z[t, T]} P_{t,x}(\zeta)$.

We will see that the perturbed test function method does not seem to work here because of a lack of smooth solutions to an equation for the perturbation. Instead we will apply the result of [14].

PROPOSITION 1.7. V^ε converges uniformly on Ω to a function W that is bounded and uniformly continuous and which is the unique viscosity solution of

$$\frac{\partial W}{\partial t} + \min_{z(\cdot) \in \mathcal{A}} W_x \cdot \left(\int_0^1 \frac{1}{f(t, y, x, z(y))} dy \right)^{-1} = 0$$

for $(t, x) \in \Omega$, and $W(T, x) = g(x)$.

Proof. Lions, Papanicolaou, and Varadhan [14] prove that V^ε converges uniformly to a bounded, continuous function W that is the viscosity solution of

$$\frac{\partial W}{\partial t} + \overline{H}(t, x, W_x) = 0,$$

where \overline{H} , the so-called *effective Hamiltonian*, is given via the solution of a *cell problem*

$$\min_{z \in Z} \{(p + \gamma'(y)) \cdot f(t, y, x, z)\} = \lambda, \quad y \in (0, 1), \quad p \in R^1,$$

for a function $\gamma \in W^{1,\infty}(R^1)$ which is 1-periodic. The pair (λ, γ) with $\lambda = \overline{H}(t, x, p)$, is unique. The problem is to compute λ .

We claim that

$$\overline{H}(t, x, p) = \min_{z(\cdot) \in \mathcal{A}} p \cdot \left(\int_0^1 \frac{1}{f(t, y, x, z(y))} dy \right)^{-1}.$$

To verify this, let $\zeta \in \mathcal{A}$ be arbitrary. Then

$$\lambda \leq (p + \gamma'(y)) \cdot f(t, y, x, \zeta(y)), \quad \text{for a.e. } y \in [0, 1],$$

so that $\lambda/f - p \leq \gamma'$. Integrating both sides on y from 0 to 1, we get

$$\lambda \int_0^1 \frac{1}{f(t, y, x, \zeta(y))} dy - p \leq \gamma(1) - \gamma(0) = 0.$$

Consequently,

$$\lambda \leq \min_{z(\cdot) \in \mathcal{A}} p \cdot \left(\int_0^1 \frac{1}{f(t, y, x, z(y))} dy \right)^{-1}.$$

For the opposite inequality, let $\zeta \in \mathcal{A}$ satisfy almost everywhere (note that ζ also depends on (t, x, p))

$$\min_{z \in Z} \{(p + \gamma'(y)) \cdot f(t, y, x, z)\} = (p + \gamma'(y)) \cdot f(t, y, x, \zeta(y)) = \lambda.$$

So, $\lambda/f - p = \gamma'$, which, by integration on y , periodicity of γ , and rearrangement, yields

$$\lambda = p \cdot \left(\int_0^1 \frac{1}{f(t, y, x, \zeta(y))} dy \right)^{-1} \geq \min_{z(\cdot) \in \mathcal{A}} p \cdot \left(\int_0^1 \frac{1}{f(t, y, x, z(y))} dy \right)^{-1}$$

and we are done. \square

The terminology in the proof seems to be due to Bensoussan, Lions, and Papanicolaou [9]. The uncontrolled analogue of this problem appears in [9, p. 8]. Note that the limit problem involves controls that depend on y . The limit optimal control problem has the dynamics

$$\frac{d\xi}{d\tau} = \left(\int_0^1 \frac{1}{f(\tau, y, \xi(\tau), \zeta(\tau, y))} dy \right)^{-1},$$

where the controls are $\zeta = \zeta(t, x)$.

The difficulty in using the perturbed test function method to prove convergence for this problem is the fact that the cell problem, while it has a Lipschitz-continuous viscosity solution, does not necessarily have a smooth classical solution.

The proposition does not seem to extend to higher dimensions.

2. The minimax problem. In this section we study the limiting behavior of the optimal control problem with dynamics (1.1), (1.2) but with cost functional given by

$$P_{t,x}^\varepsilon(\zeta) = \|h(r, \xi(r), \zeta(r))\|_{L^\infty[t,T]}.$$

The assumptions on f and h are the same as in §1, but note that here we assume that h is independent of the fast variable. The value function is defined as

$$V^\varepsilon(t, x) = \inf_{\zeta \in Z[t, T]} P_{t, x}^\varepsilon(\zeta).$$

Define the point to set map

$$Z(t, x, r) \equiv \{z \in Z \mid h(t, x, z) \leq r\},$$

and the Hamiltonian

$$H(t, s, x, r, p) \equiv \min\{p \cdot f(t, s, x, z) ; z \in Z(t, x, r)\},$$

If $Z(t, x, r)$ is empty, H is defined as $+\infty$.

The next lemma is the main result from [6] and gives the Bellman equation for the fast minimax problem.

LEMMA 2.1. *The value function V^ε is the unique continuous viscosity solution of the problem*

$$(2.1) \quad \frac{\partial V}{\partial t} + H\left(t, \frac{t}{\varepsilon}, x, V, D_x V\right) = 0 \quad \text{on } \Omega.$$

V^ε satisfies the terminal condition

$$(2.2) \quad V^\varepsilon(T, x) = \min_{z \in Z} h(T, x, z) \quad \text{on } R^n.$$

Remarks. We will also have need of the following results from [6]. It was established in [6, Prop. 4.1] that (2.1) is equivalent to the following nonlinear variational inequality on Ω

$$(2.1a) \quad \max \left\{ \frac{\partial V}{\partial t} + H\left(t, \frac{t}{\varepsilon}, x, V, D_x V\right), \min_{z \in Z} h(t, x, z) - V(t, x) \right\} = 0.$$

Equation (2.1) is an implicit obstacle problem that is made explicit in (2.1a), especially when h is independent of z . It is usually better to work with (2.1a) rather than (2.1) since (2.1a) explicitly shows that when we are dealing with supersolutions the set $Z(t, x, r)$ is not empty.

According to [6, Lemma 2.4], if $r \leq r'$ $H(t, s, x, r, p) \geq H(t, s, x, r', p)$. If $r < r'$ there exists $\varrho > 0$ depending only on r and r' such that

$$(2.3a) \quad H(\tau, s, \xi, r, p) \geq H(t, s, x, r', p) \text{ for any } (\tau, \xi) \in B_\varrho(t, x).$$

Proposition 2.5 of [6] establishes that

$$(2.3b) \quad H^*(t, s, x, r, p) = H(t, s, x, r - 0, p), \quad H_*(t, s, x, r, p) = H(t, s, x, r + 0, p).$$

Consequently, $H(t, s, x, r - 0, p)$ is upper semicontinuous and $H(t, s, x, r + 0, p)$ is lower semicontinuous in all variables. But, for fixed (t, x, r, p) , H is continuous in s .

Finally, it was shown that the essential supremum in time of h can be replaced by the supremum in the definition of V^ε . We will do so in referring to the norm in L^∞ .

Now we proceed formally to determine the limit problem in this case. We use L^p approximations to the L^∞ norm. Indeed, for fixed $\varepsilon > 0, p \geq 1$ let

$$V_p^\varepsilon(t, x) = \inf_{\zeta} \|h(r, \xi(r), \zeta(r))\|_{L^p[t, T]}.$$

Then, according to [6, Lemma 1.1], V_p^ε is the unique continuous viscosity solution of $V(T, x) = \min_{z \in Z} h(T, x, z)$ and

$$\frac{\partial V}{\partial t} + \min_{z \in Z} \left\{ D_x V \cdot f\left(t, \frac{t}{\varepsilon}, x, z\right) + \frac{1}{p} (h(t, x, z))^p V(t, x)^{1-p} \right\} = 0.$$

For $p \geq 1$ fixed, we let $\varepsilon \rightarrow 0$ and use Theorem 1.3 to obtain that $V_p = \lim_{\varepsilon \rightarrow 0} V_p^\varepsilon$ satisfies

$$\frac{\partial V}{\partial t} + \min_{z(\cdot) \in \mathcal{A}} \left\{ D_x V \cdot \bar{f}(t, x, z(\cdot)) + \frac{1}{p} V(t, x)^{1-p} \int_0^1 (h(t, x, z(s)))^p ds \right\} = 0.$$

Reasoning as in [6], by looking at the term $V^{1-p} \int h^p ds$, if a limit as $p \rightarrow \infty$ is going to exist, it must be the case that the controls $z(\cdot) \in \mathcal{A}$ must be chosen so that, for all large p

$$V_p(t, x)^p \geq \int_0^1 (h(t, x, z(s)))^p ds$$

and so

$$V_p(t, x) \geq \left(\int_0^1 (h(t, x, z(s)))^p ds \right)^{1/p},$$

which implies that

$$V(t, x) = \lim_{p \rightarrow \infty} V_p(t, x) \geq \sup_{0 \leq s \leq 1} h(t, x, z(s)).$$

Consequently, we see that the controls in the minimum must be chosen out of \mathcal{A} but that also satisfy

$$V(t, x) \geq \sup_{0 \leq s \leq 1} h(t, x, z(s)).$$

This argument can be made rigorous to yield that

$$\lim_{p \rightarrow \infty} \lim_{\varepsilon \rightarrow 0} V_p^\varepsilon(t, x) \equiv W(t, x)$$

and W satisfies, in the viscosity sense, the equation

$$\begin{aligned} & \max\{W_t + \min\{D_x W \cdot \bar{f}(t, x, z(\cdot)); z(\cdot) \in \mathcal{A}, \sup_{0 \leq s \leq 1} h(t, x, z(s)) \leq W(t, x)\}, \\ (2.4) \quad & \min_{z(\cdot) \in \mathcal{A}} \sup_{0 \leq s \leq 1} h(t, x, z(s)) - W(t, x)\} = 0, \end{aligned}$$

and the terminal condition $W(T, x) = \min_{z(\cdot) \in \mathcal{A}} \sup_{0 \leq s \leq 1} h(T, x, z(s))$. In fact, this is even true when $h = h(t, s, x, z)$, i.e., there is dependence on the fast variable. Unfortunately, it is certainly not clear that the iterated limit can be reversed (it probably cannot). Furthermore, the L^p approximations are monotone nondecreasing, actually

this is true of $(T-t)^{-p}V_p$, so that we are in fact getting a form of variational convergence ([2, Thm. 2.40]). (As an aside, for any periodic function f , $f(y/\varepsilon)$ epi-converges to $\inf_y f(y)$ and hypo-converges to $\sup_y f(y)$.) We can see that the terminal condition $V^\varepsilon(T, x) = \min_{z \in Z} h(T, T/\varepsilon, x, z)$ would cause a great deal of difficulty when there is fast variable dependence. In fact, whenever V^ε hits the obstacle $\min_{z \in Z} h(t, t/\varepsilon, x, z)$ it is not clear that there will be convergence in a strong enough sense. However, without s dependence for h , (2.4) is the correct limit equation. We will prove this fact directly using Evans' method. The case when h depends on the fast variable will be left open.

We are grateful to a reviewer for spotting a subtle, and seemingly minor error in an earlier version of this paper in which we included fast dependence for h . This point led to our realization that dependence of the obstacle on the fast variable is probably not allowed. At least we do not see how to do it.

Now we turn to the main theorem of this section.

THEOREM 2.2. *There exists a continuous function $W : [0, T] \times R^n \rightarrow R^1$ such that (at least) on a subsequence of $\{\varepsilon\}$ we have*

$$\lim_{\varepsilon \rightarrow 0} V^\varepsilon(t, x) = W(t, x).$$

The function W is the minimax value function for the optimal control problem with dynamics

$$\frac{d\xi}{d\tau} = \bar{f}(\tau, \xi(\tau), \zeta(\tau, \cdot)), \quad t < \tau \leq T, \quad \xi(t) = x,$$

and cost functional

$$\bar{P}_{t,x}(\zeta(\cdot, \cdot)) \equiv \sup_{t \leq \tau \leq T} \sup_{0 \leq s \leq 1} h(\tau, \xi(\tau), \zeta(\tau, s)).$$

where the controls $\zeta : [t, T] \times [0, 1] \rightarrow Z$ satisfy $\zeta(\tau, \cdot) \in \mathcal{Z}[t, T]$, and $\zeta(\cdot, s) \in \mathcal{A}$. Furthermore, W is the unique continuous viscosity solution of

$$(2.5) \quad \frac{\partial W}{\partial t} + \bar{H}(t, x, W, D_x W) = 0, \quad (t, x) \in [0, T] \times R^n,$$

where

$$\begin{aligned} \bar{H}(t, x, r, p) &\equiv \min\{p \cdot \bar{f}(t, x, z(\cdot)) \mid z(\cdot) \in \mathcal{A}_\infty(t, x, r)\}, \\ \mathcal{A}_\infty(t, x, r) &\equiv \{z : [0, 1] \rightarrow Z \mid \sup_{0 \leq s \leq 1} h(t, x, z(s)) \leq r\} \end{aligned}$$

\bar{H} is defined as $+\infty$ if \mathcal{A}_∞ is empty. Finally, W satisfies the terminal condition

$$W(T, x) = \min_{z \in Z} h(T, x, z).$$

Remarks. 1. It is important to distinguish the notation from §1 that $\mathcal{A} = \{z : [0, 1] \rightarrow Z \mid z \text{ is Lebesgue measurable}\}$ while in the minimax case $\mathcal{A}_\infty(t, x, r) \equiv \{z : [0, 1] \rightarrow Z \mid \sup_{0 \leq s \leq 1} h(t, x, z(s)) \leq r\}$.

2. Using the same proofs as in [6] it is established that

$$\bar{H}^*(t, x, r, p) = \bar{H}(t, x, r - 0, p), \quad \bar{H}_*(t, x, r, p) = \bar{H}(t, x, r + 0, p)$$

Also, (2.5) is equivalent to

$$(2.5a) \quad \max\left\{\frac{\partial W}{\partial t} + \overline{H}(t, x, W, D_x W), \min_{z \in Z} h(t, x, z) - W(t, x)\right\} = 0.$$

3. Let us dispense with a small technicality. Note that for each fixed $t \in [0, T]$, $x \in R^n$

$$\min_{z(\cdot) \in \mathcal{A}} \sup_{0 \leq s \leq 1} h(t, x, z(s)) = \min_{z \in Z} h(t, x, z).$$

Indeed, by taking $z(s) \equiv z$, $0 \leq s \leq 1$, where $z \in Z$ is arbitrary,

$$\min_{z(\cdot) \in \mathcal{A}} \sup_{0 \leq s \leq 1} h(t, x, z(s)) \leq \min_{z \in Z} h(t, x, z).$$

On the other hand, for any $\delta > 0$ there is a function ζ such that

$$\begin{aligned} \min_{z(\cdot) \in \mathcal{A}} \sup_{0 \leq s \leq 1} h(t, x, z(s)) &\geq h(t, x, \zeta(s, t, x)) - \delta \quad \forall 0 \leq s \leq 1 \\ &\geq \min_{z \in Z} h(t, x, z) - \delta. \end{aligned}$$

The claim is proved.

4. It is clear that $\mathcal{A}_\infty(t, x, r)$ is nonempty if and only if $\mathcal{Z}(t, x, r)$ is nonempty.

Proof of Theorem 2.2. Set

$$W^u(t, x) = \limsup_{(\varepsilon, s, y) \rightarrow (0, t, x)} V^\varepsilon(s, y)$$

and

$$W_d(t, x) = \liminf_{(\varepsilon, s, y) \rightarrow (0, t, x)} V^\varepsilon(s, y).$$

W^u is upper semicontinuous and W_d is lower semicontinuous.

We begin by showing that W^u is a subsolution of (2.5a). As usual, let $W^u - \varphi$ achieve a strict maximum at $(t_0, x_0) \in \Omega$, with φ a smooth function. If

$$(2.6) \quad \min_{z \in Z} h(t_0, x_0, z) \geq W^u(t_0, x_0)$$

then obviously,

$$\max\left\{\frac{\partial \varphi}{\partial t} + \overline{H}(t_0, x_0, W^u(t_0, x_0) - 0, D_x \varphi(t_0, x_0)), \min_{z \in Z} h(t_0, x_0, z) - W^u(t_0, x_0)\right\} \geq 0.$$

Consequently, W^u is a subsolution of (2.5a) immediately. We may therefore assume that

$$(2.7) \quad \min_{z \in Z} h(t_0, x_0, z) < W^u(t_0, x_0).$$

This implies that $\mathcal{A}_\infty(t_0, x_0, W^u(t_0, x_0) - 0)$ is not empty.

Define $\gamma : R^1 \rightarrow R^1$ as the solution of the ordinary differential equation

$$\frac{d\gamma(s)}{ds} = \overline{H}(t_0, x_0, W^u(t_0, x_0) - 0, D_x \varphi(t_0, x_0)) - H(t_0, s, x_0, W^u(t_0, x_0) - 0, D_x \varphi(t_0, x_0)),$$

if $0 \leq s \leq 1$, with γ 1-periodic. To guarantee that a periodic solution of this problem exists we need the following lemma.

LEMMA 2.2a. *For any fixed (t, x, r, p)*

$$(2.8) \quad \int_0^1 H(t, s, x, r, p) ds = \overline{H}(t, x, r, p).$$

Proof. Since $\mathcal{A}_\infty(t, x, r)$ and $Z(t, x, r)$ are either both nonempty or both empty, we may as well assume that both are nonempty and therefore that both Hamiltonians are finite. For fixed (t, x, r, p) , since $Z(t, x, r)$ is compact, we have the existence of a measurable function $\zeta(s) \equiv \zeta(t, s, x, p)$ that satisfies $\zeta \in \mathcal{A}_\infty(t, x, r)$ such that

$$\begin{aligned} \int_0^1 \min_{z \in Z(t, x, r)} p \cdot f(t, s, x, z) ds &= \int_0^1 p \cdot f(t, s, x, \zeta(s)) ds \\ &\geq \min_{z(\cdot) \in \mathcal{A}_\infty(t, x, r)} p \cdot \int_0^1 f(t, s, x, z(s)) ds. \end{aligned}$$

The other side follows from the existence for each $\delta > 0$ of a function $\zeta \in \mathcal{A}_\infty(t, x, r)$ such that

$$\begin{aligned} \min_{z(\cdot) \in \mathcal{A}_\infty(t, x, r)} p \cdot \int_0^1 f(t, s, x, z(s)) ds &\geq p \cdot \int_0^1 f(t, s, x, \zeta(s)) ds - \delta \\ &\geq \int_0^1 \min_{z \in Z(t, x, r)} p \cdot f(t, s, x, z) ds - \delta. \end{aligned}$$

So (2.8) is proved. \square

Returning to the proof of the theorem, we have that γ is a periodic function that is continuously differentiable. The derivative of γ is continuous because the constraint set in

$$H(t_0, s, x_0, W^u(t_0, x_0) - 0, D_x \varphi(t_0, x_0)) = \min_{z \in Z(t_0, x_0, W^u(t_0, x_0) - 0)} D_x \varphi(t_0, x_0) \cdot f(t_0, s, x_0, z)$$

is independent of s . Furthermore, $\gamma(0) = \gamma(1)$ by Lemma 2.1a, and, since f is periodic in s , $\gamma'(0) = \gamma'(1)$. Consequently, γ may be used as a perturbed test function. That is, for all sufficiently small $\varepsilon > 0$, $V^\varepsilon(t, x) - (\varphi + \varepsilon\gamma(t/\varepsilon))$ achieves a maximum at $(t_\varepsilon, x_\varepsilon)$ and, by Lemma A.3, $(t_\varepsilon, x_\varepsilon) \rightarrow (t_0, x_0)$, and $V^\varepsilon(t_\varepsilon, x_\varepsilon) \rightarrow W^u(t_0, x_0)$ as $\varepsilon \rightarrow 0$. Using (2.7) and condition (A), it is easy to see that for all sufficiently small ε ,

$$(2.9) \quad \min_{z \in Z} h(t_\varepsilon, x_\varepsilon, z) < V^\varepsilon(t_\varepsilon, x_\varepsilon).$$

Of course this implies that $Z(t_\varepsilon, x_\varepsilon, V^\varepsilon(t_\varepsilon, x_\varepsilon))$ is nonempty for all small ε . Therefore, using the fact that V^ε is a (continuous) subsolution of (2.1), we conclude that

$$\begin{aligned} 0 &\leq \varphi_t(t_\varepsilon, x_\varepsilon) + \frac{d\gamma(\frac{t_\varepsilon}{\varepsilon})}{ds} + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, V^\varepsilon(t_\varepsilon, x_\varepsilon) - 0, D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \\ &= \varphi_t(t_\varepsilon, x_\varepsilon) + \overline{H}(t_0, x_0, W^u(t_0, x_0) - 0, D_x \varphi(t_0, x_0)) \\ &\quad - H\left(t_0, \frac{t_\varepsilon}{\varepsilon}, x_0, W^u(t_0, x_0) - 0, D_x \varphi(t_0, x_0)\right) \\ &\quad + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, V^\varepsilon(t_\varepsilon, x_\varepsilon) - 0, D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \\ &\leq \varphi_t(t_0, x_0) + \overline{H}(t_0, x_0, W^u(t_0, x_0) - 0, D_x \varphi(t_0, x_0)) + o_\varepsilon(1), \end{aligned}$$

as $\varepsilon \rightarrow 0$. We have used here the upper semicontinuity in all variables of $H(t, s, x, r + 0, p)$, the smoothness of φ and the facts $(t_\varepsilon, x_\varepsilon) \rightarrow (t_0, x_0)$, $V^\varepsilon(t_\varepsilon, x_\varepsilon) \rightarrow W^u(t_0, x_0)$. The conclusion, using (2.7), is that W^u is a subsolution of (2.5).

Next, we will prove that W_d is a supersolution. The proof is very similar to the subsolution case.

Let $W_d - \varphi$ achieve a strict minimum at (t_0, x_0) . For any $\varepsilon > 0$ let $V^\varepsilon - \varphi$ achieve a minimum at $(t_\varepsilon, x_\varepsilon)$ with $(t_\varepsilon, x_\varepsilon) \rightarrow (t_0, x_0)$, $V^\varepsilon(t_\varepsilon, x_\varepsilon) \rightarrow W_d(t_0, x_0)$. Since V^ε is a supersolution of (2.1), or equivalently, (2.1a), we have

$$(2.10) \quad V^\varepsilon(t_\varepsilon, x_\varepsilon) \geq \min_{z \in Z} h(t_\varepsilon, x_\varepsilon, z).$$

From (2.10) and the continuity of h , letting $\varepsilon \rightarrow 0$, we get

$$(2.11) \quad W_d(t_0, x_0) \geq \min_{z \in Z} h(t_0, x_0, z).$$

This implies that $\mathcal{A}_\infty(t_0, x_0, W_d(t_0, x_0) + 0)$ is not empty.

Define $\gamma: R^1 \rightarrow R^1$ as the solution of the ordinary differential equation

$$(2.12) \quad \frac{d\gamma(s)}{ds} = \overline{H}(t_0, x_0, W_d(t_0, x_0) + 0, D_x \varphi(t_0, x_0)) - H(t_0, s, x_0, W_d(t_0, x_0) + 0, D_x \varphi(t_0, x_0)),$$

if $0 \leq s \leq 1$, with γ 1-periodic. Again, a periodic solution is guaranteed to exist by Lemma 2.1a. Furthermore, since $H(t, s, x, r + 0, p)$ is continuous in s , $d\gamma/ds$ will be continuous, and $\gamma(0) = \gamma(1)$, $\gamma'(0) = \gamma'(1)$. Note also that $H(t_0, s, x_0, W_d(t_0, x_0) + 0, D_x \varphi(t_0, x_0))$ is finite because of (2.11).

Again using Lemma A.3, expressed for supersolutions, for all $\varepsilon > 0$ sufficiently small, the function $V^\varepsilon(t, x) - (\varphi(t, x) + \varepsilon \gamma(\frac{t}{\varepsilon}))$ achieves a minimum at $(t_\varepsilon, x_\varepsilon)$, $(t_\varepsilon, x_\varepsilon) \rightarrow (t_0, x_0)$ and $V^\varepsilon(t_\varepsilon, x_\varepsilon) \rightarrow W_d(t_0, x_0)$ as $\varepsilon \rightarrow 0$. Since V^ε is a continuous supersolution of (2.1) we obtain from (2.12) evaluated at $s = t_\varepsilon/\varepsilon$ and the fact that (2.10) implies that $Z(t_\varepsilon, x_\varepsilon, V^\varepsilon(t_\varepsilon, x_\varepsilon) + 0)$ is nonempty,

$$\begin{aligned} 0 &\geq \varphi_t(t_\varepsilon, x_\varepsilon) + \frac{d\gamma(\frac{t_\varepsilon}{\varepsilon})}{ds} + H_*\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, V^\varepsilon(t_\varepsilon, x_\varepsilon), D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \\ &= \varphi_t(t_\varepsilon, x_\varepsilon) + \frac{d\gamma(\frac{t_\varepsilon}{\varepsilon})}{ds} + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, V^\varepsilon(t_\varepsilon, x_\varepsilon) + 0, D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \\ &= \varphi_t(t_\varepsilon, x_\varepsilon) + \overline{H}(t_0, x_0, W_d(t_0, x_0) + 0, D_x \varphi(t_0, x_0)) \\ &\quad - H\left(t_0, \frac{t_\varepsilon}{\varepsilon}, x_0, W_d(t_0, x_0) + 0, D_x \varphi(t_0, x_0)\right) \\ &\quad + H\left(t_\varepsilon, \frac{t_\varepsilon}{\varepsilon}, x_\varepsilon, V^\varepsilon(t_\varepsilon, x_\varepsilon) + 0, D_x \varphi(t_\varepsilon, x_\varepsilon)\right) \end{aligned}$$

By letting $\varepsilon \rightarrow 0$ and using the lower semicontinuity of $H(t, s, x, r + 0, p)$, we conclude that

$$\varphi_t(t_0, x_0) + \overline{H}(t_0, x_0, W_d(t_0, x_0) + 0, D_x \varphi(t_0, x_0)) \leq 0.$$

Therefore, W_d is a supersolution of (2.5).

Since $V^\varepsilon(T, x) = \min_{z \in Z} h(T, x, z)$, it is an immediate consequence of the definitions of W^u and W_d and continuity of h that

$$W^u(T, x) = W_d(T, x) = \min_{z \in Z} h(T, x, z).$$

We have shown that W^u is a subsolution and W_d is a supersolution of (2.5), both satisfying the same terminal condition. If we can establish that $W^u \leq W_d$ then convergence to a continuous function will be shown. However, this fact is a comparison principle on which uniqueness of solutions of (2.5) is based. The statements we need are in the next theorem, the first part of which is essentially Theorem 4.2 in [6].

THEOREM 2.3. a) Let $u : [0, T] \times R^n \rightarrow R^1$ ($v : [0, T] \times R^n \rightarrow R^1$) be an upper (lower) semicontinuous subsolution (supersolution) of (2.5) both satisfying the terminal condition $u(T, x) = v(T, x) = \min_{z \in Z} h(T, x, z)$. Then, $u \leq v$ on $[0, T] \times R^n$.

b) The unique continuous viscosity solution of (2.5) is given by

$$W(t, x) = \inf_{\zeta(\cdot, \cdot)} \sup_{t \leq \tau \leq T, 0 \leq s \leq 1} h(\tau, \xi(\tau), \zeta(\tau, s)),$$

where

$$\frac{d\xi}{d\tau} = \int_0^1 f(\tau, s, \xi(\tau), \zeta(\tau, s)), \quad \xi(t) = x.$$

Assuming that Theorem 2.3 is true, let us complete the proof of Theorem 2.1. We have shown that W^u is an upper semicontinuous subsolution and W_d is a lower semicontinuous supersolution, which agree at time $t = T$. By Theorem 2.3, $W^u \leq W_d$, but then, $W(t, x) \equiv W^u(t, x) = W_d(t, x)$ is a continuous viscosity solution of (2.5) and we are done.

Proof of Theorem 2.3(a). We will only sketch the proof of the theorem since it is very similar to [6, Thm. 4.2]. Refer to that proof for complete details.

First, we may assume that $u(t, x) > \min_{z \in Z} h(t, x, z)$ because otherwise, since v is a supersolution, $v(t, x) \geq \min_{z \in Z} h(t, x, z) \geq u(t, x)$ and there is nothing to prove. Set

$$v'(t, x) = v(t, x) + \frac{\beta}{t},$$

where $\beta > 0$. Then $v' \geq v$ and it is easy to check that v' is a lower semicontinuous viscosity supersolution of

$$v'_t + \bar{H}(t, x, v', D_x v') + \frac{\beta}{t^2} = 0.$$

Set $w(t, x, y) = u(t, x) - v'(t, y)$. Since u is an upper semicontinuous subsolution, it is immediate that w is an upper semicontinuous subsolution of the equation

$$w_t + \bar{H}^*(t, x, w + v'(t, y), D_x w) - \bar{H}_*(t, y, -w + u(t, x), -D_y w) - \frac{\beta}{t^2} \geq 0.$$

Now consider $M_\alpha \equiv \sup_{[0, T] \times R^{2n}} (w(t, x, y) - \alpha/2 |x - y|^2)$, for $\alpha > 0$. Then $M_\alpha < \infty$ for large α and if $(t_\alpha, x_\alpha, y_\alpha)$ satisfies

$$\lim_{\alpha \rightarrow \infty} (M_\alpha - w(t_\alpha, x_\alpha, y_\alpha) + \frac{\alpha}{2} |x_\alpha - y_\alpha|^2) = 0,$$

then (i) $\alpha|x_\alpha - y_\alpha|^2 \rightarrow 0$, and (ii) $M_\alpha \rightarrow u(t', x') - v'(t', y') = \max(u - v')$.

Assume that $\max(u - v') > 0$. Then $0 < t' < T$ and from the definition of viscosity subsolution, with the test function $\alpha/2|x - y|^2$ we have that

$$\begin{aligned} & \bar{H}^*(t_\alpha, x_\alpha, w(t_\alpha, x_\alpha, y_\alpha) + v'(t_\alpha, y_\alpha), \alpha(x_\alpha - y_\alpha)) \\ & - \bar{H}_*(t_\alpha, y_\alpha, -w(t_\alpha, x_\alpha, y_\alpha) + u(t_\alpha, x_\alpha), \alpha(x_\alpha - y_\alpha)) - \frac{\beta}{t^2} \geq 0. \end{aligned}$$

Since we may assume that $w(t_\alpha, x_\alpha, y_\alpha) > 0$ for large α , we have that

$$u(t_\alpha, x_\alpha) = w(t_\alpha, x_\alpha, y_\alpha) + v'(t_\alpha, y_\alpha) > v'(t_\alpha, y_\alpha).$$

Using the formulas for the upper and lower semicontinuous envelopes of \bar{H} and monotonicity properties of $\bar{H}(\cdot, \cdot, r, \cdot)$ we obtain

$$\bar{H}^*(t_\alpha, x_\alpha, u(t_\alpha, x_\alpha), \alpha(x_\alpha - y_\alpha)) = \bar{H}(t_\alpha, x_\alpha, u(t_\alpha, x_\alpha) - 0, \alpha(x_\alpha - y_\alpha))$$

and, for large enough α ,

$$\begin{aligned} \bar{H}^*(t_\alpha, y_\alpha, v'(t_\alpha, y_\alpha), \alpha(x_\alpha - y_\alpha)) &= \bar{H}(t_\alpha, y_\alpha, v'(t_\alpha, y_\alpha) + 0, \alpha(x_\alpha - y_\alpha)) \\ &\geq \bar{H}(t_\alpha, y_\alpha, u(t_\alpha, x_\alpha) - 0, \alpha(x_\alpha - y_\alpha)). \end{aligned}$$

Consequently,

$$\begin{aligned} 0 &\leq \bar{H}(t_\alpha, x_\alpha, u(t_\alpha, x_\alpha) - 0, \alpha(x_\alpha - y_\alpha)) - \bar{H}(t_\alpha, y_\alpha, v'(t_\alpha, y_\alpha) + 0, \alpha(x_\alpha - y_\alpha)) - \frac{\beta}{t_\alpha^2} \\ &\leq \bar{H}(t_\alpha, x_\alpha, u(t_\alpha, x_\alpha) - 0, \alpha(x_\alpha - y_\alpha)) - \bar{H}(t_\alpha, y_\alpha, u(t_\alpha, x_\alpha) - 0, \alpha(x_\alpha - y_\alpha)) - \frac{\beta}{t_\alpha^2} \\ &\leq \bar{H}(t', y_\alpha, u(t_\alpha, x_\alpha) - 0, \alpha(x_\alpha - y_\alpha)) \\ &\quad - \bar{H}(t', y_\alpha, u(t_\alpha, x_\alpha) - 0, \alpha(x_\alpha - y_\alpha)) - \frac{\beta}{t'^2} + o_{1/\alpha}(1) \end{aligned}$$

where we have used the analogue of (2.3a) for \bar{H} in the last line. Therefore, $-\beta/t'^2 \geq 0$ which is clearly a contradiction. Since $\beta > 0$ was arbitrary, we conclude that $u \leq v$.

Details which remain consist of guaranteeing that the maxima in the proof are actually achieved. This is done by penalizing large x values (see [6]).

For the proof of part (b) we need only state that the value function W in the statement of (b) is the, unique viscosity solution of (2.5) satisfying the terminal condition $W(T, x) = \min_{z \in Z} h(T, x, z)$. The proof of this is virtually identical to [6, Thm. 4.2]. Therefore, the limit of the fast minimax problem must be this solution, and thus the value function of (b). \square

Remark. We used the uniqueness theorem to establish convergence of V^ε to a continuous function W . When h is independent of the fast variable this is not necessary since uniform convergence can be established in a way similar to Lemma 1.2. When h is independent of t/ε , the proof of [6, Prop. 1.5] along with (A) shows that $\|V^\varepsilon\| \leq K$ and $\|DV^\varepsilon\| \leq K$ in the uniform norm, with K independent of ε . Therefore, V^ε has a uniformly convergent subsequence to a Lipschitz continuous function W . The proof of [6, Prop. 1.5] is a direct proof and does not involve second-order partial differential equation approximations. When h depends on t/ε , we should *not* expect uniform convergence.

We conclude this paper by stating the result for the minimax problem when the controls are inertial. Again, assume that $Z = [0, 1]^p$ and the dynamics f and h are 1-periodic in the z variable.

THEOREM 2.3. Let $V^{\varepsilon,L}(t, x, z)$ denote the value function for the fast minimax problem when the controls ζ must be chosen from $Z_z^L[t, T]$. Then

$$U(t, x) = \lim_{L \rightarrow \infty} \lim_{\varepsilon \rightarrow 0} V^{\varepsilon,L}(t, x, z)$$

is the unique continuous viscosity solution of

$$\max \left\{ U_t + \min \left\{ D_x U \cdot \int_0^1 f(t, s, x, z) ds : z \in Z(t, x, U(t, x)) \right\}, \right. \\ \left. \min_{z \in Z} h(t, x, z) - U(t, x) \right\} = 0,$$

on Ω and $U(T, x) = \min_{z \in Z} h(T, x, z)$. Further,

$$U(t, x) = \inf_{\zeta \in Z[t, T]} \sup_{t \leq \tau \leq T} h(\tau, \xi(\tau), \zeta(\tau)),$$

where

$$\frac{d\xi}{d\tau} = \int_0^1 f(\tau, s, \xi(\tau), \zeta(\tau)) d\tau, \quad t < \tau \leq T, \quad \xi(t) = x.$$

Finally, $W(t, x) = \lim_{\varepsilon \rightarrow 0} \lim_{L \rightarrow \infty} V^{\varepsilon,L}(t, x, z)$, where W is given in Theorem 2.2, and $W(t, x) \leq U(t, x)$ on $\bar{\Omega}$.

The proof of this theorem is similar to that of Theorem 1.5 if we use the results concerning Lipschitz controls for minimax control problems from [4]. Once again we see that Lipschitz controls cannot keep up with rapid oscillations.

Remark. The results of this section are true for the more general fast minimax cost

$$V^{\varepsilon}(t, x) = \inf_{\zeta \in Z[t, T]} \sup_{t \leq \tau \leq T} \left\{ \int_t^{\tau} k\left(r, \frac{r}{\varepsilon}, \xi(r), \zeta(r)\right) dr + h(\tau, \xi(\tau), \zeta(\tau)) \right\},$$

where k is a given function satisfying a condition like (A). In this case $W(t, x) = \lim_{\varepsilon \rightarrow 0} V^{\varepsilon}(t, x)$ satisfies

$$W_t + \min_{z(\cdot) \in \mathcal{A}_{\infty}(t, x, W(t, x))} \{D_x W \cdot \bar{f}(t, x, z(\cdot)) + \bar{k}(t, x, z(\cdot))\} = 0,$$

where $\bar{k}(t, x, z(\cdot)) = \int_0^1 k(t, s, x, z(s)) ds$.

Acknowledgments. I am grateful to the reviewers for their careful reading of this paper. One comment led to a substantial simplification of the proofs and another comment led to a revision of Theorem 2.1. I also thank L. C. Evans for sending me a preprint of [14].

REFERENCES

- [1] G. ARONSSON, *Minimization problems for the functional* $\sup_x F(x, f(x), f'(x))$, Ark. Mat., 6 (1965), pp. 33–53.
- [2] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, London, 1983.

- [3] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, Math. Modelling Numer. Anal., 21 (1987), pp. 557–579.
- [4] E. N. BARRON, *Differential games with maximum cost*, Nonlinear Anal., TMA 14 (1990), pp. 971–989.
- [5] ———, *Differential games with Lipschitz control functions and fixed initial control positions*, J. Differential Equations, 26 (1977), pp. 161–180.
- [6] E. N. BARRON AND H. ISHII, *The Bellman equation for minimizing the maximum cost*, Nonlinear Anal., TMA 13 (1989), pp. 1067–1090.
- [7] E. N. BARRON, L. C. EVANS, AND R. JENSEN, *Viscosity solutions of Isaacs' equation and differential games with Lipschitz controls*, J. Differential Equations, 53 (1984), pp. 213–233.
- [8] E. N. BARRON AND R. JENSEN, *Semicontinuous viscosity solutions for Hamilton–Jacobi equations with convex Hamiltonian*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.
- [9] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North Holland, New York, 1978.
- [10] R. BULIRSCH, F. MONTRONE, AND H. J. PESCH, *Abort landing in the presence of windshear as a minimax optimal control problem, Part 1: Necessary conditions*, J. Optim. Theory and Appl., 70 (1991), pp. 1–23.
- [11] F. CHAPLAIS, *Averaging and deterministic optimal control*, SIAM J. Control Optim., 25 (1987), pp. 767–780.
- [12] M. G. CRANDALL, L. C. EVANS, AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
- [13] L. C. EVANS, *The perturbed test function method for viscosity solutions of nonlinear pde*, Proc. Royal Soc. Edinburgh, 111A (1989), pp. 359–375.
- [14] P. L. LIONS, G. PAPANICOLAOU, AND S. R. S. VARADHAN, *Homogenization of Hamilton Jacobi equations*, to appear.
- [15] S. PENG, *Analyse Asymptotique et probleme homogeneise en controle optimal avec vibrations rapides*, SIAM J. Control Optim., 27 (1989), pp. 673–696.